```
In [1]: #Aline Barboza Soares RA:1800136  E-mail: aline.soares@aluno.faculdadei
        mpacta.com.br
        #Luciana silveira     RA:1800005  E-mail: luciana.silveira@aluno.faculd
        adeimpacta.com.br
```

```
In [2]: #Leia a base de dados de nome filmes.csv
```

```
In [1]: import pandas
```

```
In [2]: dados = pandas.read_csv("D:/DP-FACULDADE/filmes.csv")
```

```
In [3]: type(dados)
```

```
Out[3]: pandas.core.frame.DataFrame
```

```
In [4]: #Exiba os dados brutos no Jupyter Notebook
```

```
In [5]: print(dados)
```

```
      color      director_name  num_critic_for_reviews  duration  \
0     Color      James Cameron                   723.0     178.0   
1     Color     Gore Verbinski                   302.0     169.0   
2     Color         Sam Mendes                   602.0     148.0   
3     Color  Christopher Nolan                   813.0     164.0   
4       NaN        Doug Walker                     NaN       NaN   
...     ...                ...                     ...       ...   
5038  Color        Scott Smith                     1.0      87.0   
5039  Color                NaN                    43.0      43.0   
5040  Color   Benjamin Roberds                    13.0      76.0   
5041  Color        Daniel Hsia                    14.0     100.0   
5042  Color           Jon Gunn                    43.0      90.0   

      director_facebook_likes  actor_3_facebook_likes      actor_2_name
  \
0                         0.0                       855.0  Joel David Moore

1                       563.0                      1000.0     Orlando Bloom

2                         0.0                       161.0      Rory Kinnear

3                     22000.0                     23000.0    Christian Bale

4                       131.0                         NaN        Rob Walker

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

5038                      2.0                       318.0     Daphne Zuniga

5039                      NaN                       319.0     Valorie Curry

5040                      0.0                         0.0     Maxwell Moody

5041                      0.0                       489.0     Daniel Henney

5042                     16.0                        16.0  Brian Herzlinger


      actor_1_facebook_likes        gross                             gen
res  \
0                     1000.0  760505847.0  Action|Adventure|Fantasy|Sci
-Fi
1                    40000.0  309404152.0          Action|Adventure|Fant
asy
2                    11000.0  200074175.0         Action|Adventure|Thril
ler
3                    27000.0  448130642.0                   Action|Thril
ler
4                      131.0          NaN                       Document
ary
...                      ...          ...
...
5038                   637.0          NaN                      Comedy|Dr
ama
```

```
5039                    841.0         NaN     Crime|Drama|Mystery|Thril
ler
5040                      0.0         NaN             Drama|Horror|Thril
ler
5041                    946.0     10443.0             Comedy|Drama|Roma
nce
5042                     86.0     85222.0                      Document
ary

      ... num_user_for_reviews language  country  content_rating
budget  \
0     ...               3054.0  English      USA           PG-13  23700
0000.0
1     ...               1238.0  English      USA           PG-13  30000
0000.0
2     ...                994.0  English       UK           PG-13  24500
0000.0
3     ...               2701.0  English      USA           PG-13  25000
0000.0
4     ...                  NaN      NaN      NaN             NaN
   NaN
...   ...                  ...      ...      ...             ...
   ...
5038  ...                  6.0  English   Canada             NaN
   NaN
5039  ...                359.0  English      USA           TV-14
   NaN
5040  ...                  3.0  English      USA             NaN
1400.0
5041  ...                  9.0  English      USA           PG-13
   NaN
5042  ...                 84.0  English      USA              PG
1100.0

      title_year actor_2_facebook_likes imdb_score  aspect_ratio  \
0         2009.0                  936.0        7.9          1.78
1         2007.0                 5000.0        7.1          2.35
2         2015.0                  393.0        6.8          2.35
3         2012.0                23000.0        8.5          2.35
```

```
4           NaN                     12.0        7.1           NaN
...         ...                      ...        ...           ...
5038     2013.0                    470.0        7.7           NaN
5039        NaN                    593.0        7.5         16.00
5040     2013.0                      0.0        6.3           NaN
5041     2012.0                    719.0        6.3          2.35
5042     2004.0                     23.0        6.6          1.85

     movie_facebook_likes
0                   33000
1                       0
2                   85000
3                  164000
4                       0
...                   ...
5038                   84
5039                32000
5040                   16
5041                  660
5042                  456

[5043 rows x 28 columns]
```

In [6]:
```
#3. Realize as seguintes tarefas de tratamentos nas informações:
#Coluna do tipo String, caso esteja vazio, inserir NA
```

In [7]:
```
dados.isnull().sum()
```

Out[7]:
```
color                         19
director_name                104
num_critic_for_reviews        50
duration                      15
director_facebook_likes      104
actor_3_facebook_likes        23
actor_2_name                  13
actor_1_facebook_likes         7
gross                        884
genres                         0
actor_1_name                   7
```

```
movie_title                    0
num_voted_users                0
cast_total_facebook_likes      0
actor_3_name                  23
facenumber_in_poster          13
plot_keywords                153
movie_imdb_link                0
num_user_for_reviews          21
language                      12
country                        5
content_rating               303
budget                       492
title_year                   108
actor_2_facebook_likes        13
imdb_score                     0
aspect_ratio                 329
movie_facebook_likes           0
dtype: int64
```

In [8]:
```
dados.director_name.fillna('NA', inplace=True)
```

In [9]:
```
dados.color.fillna('NA', inplace=True)
dados.actor_2_name.fillna('NA', inplace=True)
```

In [10]:
```
dados.genres.fillna('NA', inplace=True)
dados.language.fillna('NA', inplace=True)
dados.country.fillna('NA', inplace=True)
dados.actor_2_name.fillna('NA', inplace=True)
```

In [11]:
```
dados.content_rating.fillna('NA', inplace=True)
```

In [12]:
```
#Coluna do tipo numérica, caso esteja vazio, inserir o número 0.
```

In [13]:
```
dados.duration.fillna(0, inplace=True)
```

In [14]:
```
dados.num_critic_for_reviews.fillna(0, inplace=True)
dados.director_facebook_likes.fillna(0, inplace=True)
dados.actor_3_facebook_likes.fillna(0, inplace=True)
dados.actor_1_facebook_likes.fillna(0, inplace=True)
dados.gross.fillna(0, inplace=True)
dados.num_user_for_reviews.fillna(0, inplace=True)
dados.budget.fillna(0, inplace=True)
dados.title_year.fillna(0, inplace=True)
dados.actor_2_facebook_likes.fillna(0, inplace=True)
dados.imdb_score.fillna(0, inplace=True)
dados.aspect_ratio.fillna(0, inplace=True)
dados.movie_facebook_likes.fillna(0, inplace=True)
```

In [15]:
```
#Todas as letras devem estar maiúsculas;
```

In [16]:
```
dados['director_name_up'] = dados.director_name.map(lambda x:x.upper())
dados['color_up'] = dados.color.map(lambda x:x.upper())
dados['actor_2_name_up'] = dados.actor_2_name.map(lambda x:x.upper())
dados['genres_up'] = dados.genres.map(lambda x:x.upper())
dados['language_up'] = dados.language.map(lambda x:x.upper())
dados['country_up'] = dados.country.map(lambda x:x.upper())
```

In [17]:
```
#Exiba novamente o dataframe com todos os tratamentos aplicados.
```

In [18]:
```
print(dados)
```

```
      color      director_name  num_critic_for_reviews  duration  \
0     Color      James Cameron                   723.0     178.0
1     Color     Gore Verbinski                   302.0     169.0
2     Color         Sam Mendes                   602.0     148.0
3     Color  Christopher Nolan                   813.0     164.0
4        NA        Doug Walker                     0.0       0.0
...     ...                ...                     ...       ...
5038  Color        Scott Smith                     1.0      87.0
5039  Color                 NA                    43.0      43.0
5040  Color   Benjamin Roberds                    13.0      76.0
5041  Color        Daniel Hsia                    14.0     100.0
5042  Color           Jon Gunn                    43.0      90.0
```

```
      director_facebook_likes  actor_3_facebook_likes      actor_2_name
  \
0                         0.0                   855.0  Joel David Moore

1                       563.0                  1000.0     Orlando Bloom

2                         0.0                   161.0      Rory Kinnear

3                     22000.0                 23000.0    Christian Bale

4                       131.0                     0.0        Rob Walker

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

5038                      2.0                   318.0     Daphne Zuniga

5039                      0.0                   319.0     Valorie Curry

5040                      0.0                     0.0     Maxwell Moody

5041                      0.0                   489.0     Daniel Henney

5042                     16.0                    16.0  Brian Herzlinger


      actor_1_facebook_likes        gross                          gen
res  \
0                     1000.0  760505847.0  Action|Adventure|Fantasy|Sci
-Fi
1                    40000.0  309404152.0         Action|Adventure|Fant
asy
2                    11000.0  200074175.0        Action|Adventure|Thril
ler
3                    27000.0  448130642.0                   Action|Thril
ler
4                      131.0          0.0                       Document
ary
...                      ...          ...
...
```

```
5038                   637.0          0.0                     Comedy|Dr
ama
5039                   841.0          0.0     Crime|Drama|Mystery|Thril
ler
5040                     0.0          0.0            Drama|Horror|Thril
ler
5041                   946.0      10443.0             Comedy|Drama|Roma
nce
5042                    86.0      85222.0                      Document
ary

      ... actor_2_facebook_likes imdb_score  aspect_ratio  \
0     ...                  936.0        7.9          1.78
1     ...                 5000.0        7.1          2.35
2     ...                  393.0        6.8          2.35
3     ...                23000.0        8.5          2.35
4     ...                   12.0        7.1          0.00
...   ...                    ...        ...           ...
5038  ...                  470.0        7.7          0.00
5039  ...                  593.0        7.5         16.00
5040  ...                    0.0        6.3          0.00
5041  ...                  719.0        6.3          2.35
5042  ...                   23.0        6.6          1.85

      movie_facebook_likes   director_name_up  color_up   actor_2_name_
up  \
0                    33000      JAMES CAMERON     COLOR  JOEL DAVID MOO
RE
1                        0     GORE VERBINSKI     COLOR      ORLANDO BLO
OM
2                    85000         SAM MENDES     COLOR      RORY KINNE
AR
3                   164000  CHRISTOPHER NOLAN     COLOR     CHRISTIAN BA
LE
4                        0        DOUG WALKER        NA         ROB WALK
ER
...                    ...                ...       ...
...
5038                    84        SCOTT SMITH     COLOR      DAPHNE ZUNI
```

```
GA
5039                  32000                  NA     COLOR     VALORIE CUR
RY
5040                     16    BENJAMIN ROBERDS     COLOR     MAXWELL MOO
DY
5041                    660         DANIEL HSIA     COLOR     DANIEL HENN
EY
5042                    456            JON GUNN     COLOR  BRIAN HERZLING
ER

                            genres_up  language_up country_up
0     ACTION|ADVENTURE|FANTASY|SCI-FI      ENGLISH        USA
1            ACTION|ADVENTURE|FANTASY      ENGLISH        USA
2           ACTION|ADVENTURE|THRILLER      ENGLISH         UK
3                     ACTION|THRILLER      ENGLISH        USA
4                         DOCUMENTARY           NA         NA
...                               ...          ...        ...
5038                     COMEDY|DRAMA      ENGLISH     CANADA
5039      CRIME|DRAMA|MYSTERY|THRILLER      ENGLISH        USA
5040            DRAMA|HORROR|THRILLER      ENGLISH        USA
5041             COMEDY|DRAMA|ROMANCE      ENGLISH        USA
5042                      DOCUMENTARY      ENGLISH        USA

[5043 rows x 34 columns]
```

In [19]:

```
#Construa um gráfico com as colunas que você quiser da base.
```

In [20]:

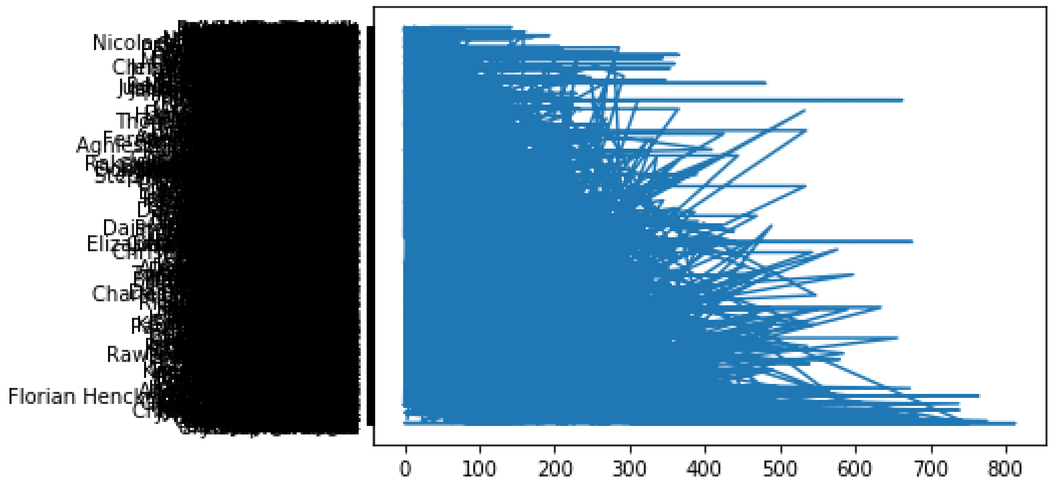
```
import matplotlib.pyplot
```

In [21]:

```
matplotlib.pyplot.plot(dados.num_critic_for_reviews, dados.director_nam
e)
```

Out[21]:

```
[<matplotlib.lines.Line2D at 0x26d483f3f08>]
```



In [22]:
```
matplotlib.pyplot.show()
```

In [23]:
```
#5.Aplicar o K-Médias para a base já tratada;
    #Escolha duas colunas numéricas na base de dados e identifique uma
 como sendo
    #X e a outra Y e faça a visualização da informação pelo gráfico sca
tter.
```

In [34]:
```
from pandas import DataFrame
```

In [35]:
```
import matplotlib.pyplot as plt
```

In [36]:
```
from sklearn.cluster import KMeans
%matplotlib inline
```

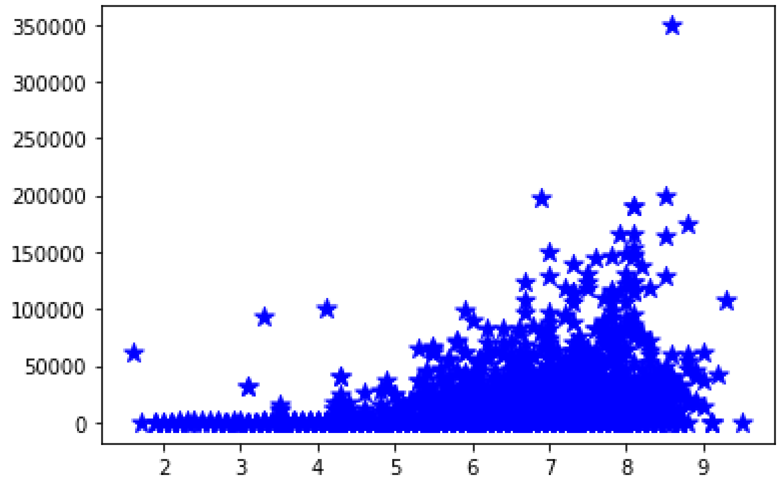
In [37]:
```
Data = {'x': dados.imdb_score,
        'y': dados.movie_facebook_likes
       }
```

In [38]:
```
df = DataFrame(Data,columns=['x','y'])
```

In [39]:
```
#Escolha duas colunas numéricas na base de dados e identifique uma como
 sendo
#X e a outra Y e faça a visualização da informação pelo gráfico scatte
r.
```

In [40]:
```
plt.scatter(Data['x'],Data['y'], label = 'Pontos', color = 'b', marker
= '*', s = 100)
```

Out[40]:
```
<matplotlib.collections.PathCollection at 0x26d4eb4a248>
```



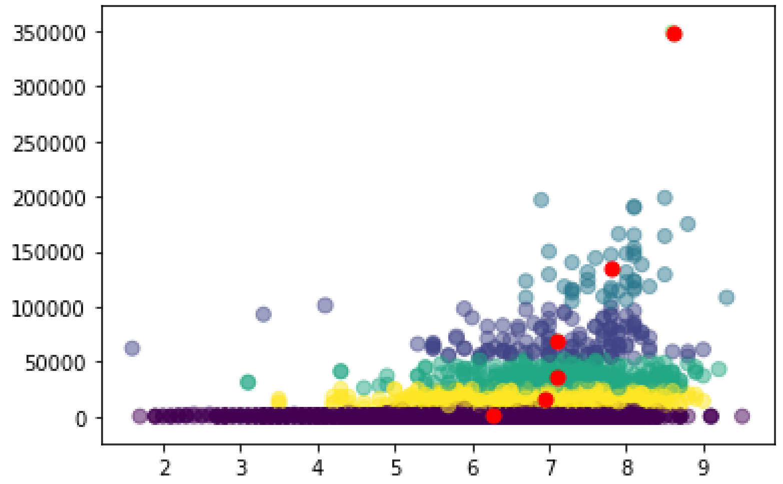
In [41]:
```
#Aplique o K-Médias para esse conjunto de dados, exibindo o gráfico com
 os seus
#centroides em vermelho e os clusters obtidos.
```

In [42]:
```
kmeans = KMeans(n_clusters=6).fit(df)
centroids = kmeans.cluster_centers_
print(centroids)
```

```
[[6.26170543e+00 3.44557881e+02]
 [7.09007634e+00 6.94274809e+04]
 [7.81162791e+00 1.35139535e+05]
 [7.10993151e+00 3.59486301e+04]
 [8.60000000e+00 3.49000000e+05]
 [6.94830028e+00 1.53937677e+04]]
```

In [43]:
```
plt.scatter(df['x'], df['y'], c= kmeans.labels_.astype(float), s=50, al
pha=0.5)
plt.scatter(centroids[:, 0], centroids[:, 1], c='red', s=50)
```

Out[43]:
```
<matplotlib.collections.PathCollection at 0x26d4ebb7d48>
```



In [44]:
```
#Justifique o porquê da escolha desta quantidade de clusters no seu mod
elo.
```

In [102]:
```
#Para calcular a quantidade de clusters devemos levar em consideração u
m numero que torne a distancia entre
#clusters seja a menor possível.
```

In [ ]:
```

```