Data Analysis of IMDB movies data

1. Read Data

In [57]:
```
# Read data from .csv file

import numpy as np
import pandas as pd

data = pd.read_csv(r'/Users/babarhussain/Documents/movies.csv')
data
```

Out[57]:

	Rank	Title	Genre	Description	Director	
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chr Die Coo
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noo Loga Green, M
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James M Taylor-J
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	McConau Wither
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will S Leto, Mar
...	...	...	...	...	...	
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiv Nicole K

996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Laur Heather
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robe Br Cassi
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Ada Mil Midd
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Ke Jenr Robbie

1000 rows × 12 columns

In [7]:
```
len(data)
```

Out[7]: 1000

In [9]:
```
len(data.columns)
```

Out[9]: 12

head( )

In [12]:
```
# Returns the top 5 rows in the dataset by default

data.head()
```

Out[12]:

	Rank	Title	Genre	Description	Director	
0	1	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pı Diesel, Cooper,
1	2	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi F Logan M Green, Mich
2	3	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvc Taylor-Joy, H F
3	4	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	M McConaughe Witherspoc
4	5	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smitl Leto, Margot V

tail( )

In [15]:
```
# Returns the bottom 5 rows in the dataset by default
data.tail()
```

Out[15]:

	Rank	Title	Genre	Description	Director	Actors
995	996	Secret in Their Eyes	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor Nicole Kidman Julia Roberts..
996	997	Hostel: Part II	Horror	Three American college students studying abroa...	Eli Roth	Lauren German Heathe Matarazzo, Bijou Philli..
997	998	Step Up 2: The Streets	Drama,Music,Romance	Romantic sparks occur between two dance studen...	Jon M. Chu	Robert Hoffman Briana Evigan Cassie Ventura,..
998	999	Search Party	Adventure,Comedy	A pair of friends embark on a mission to reuni...	Scot Armstrong	Adam Pally, T.J Miller, Thomas Middleditch,Sh..
999	1000	Nine Lives	Comedy,Family,Fantasy	A stuffy businessman finds himself trapped ins...	Barry Sonnenfeld	Kevin Spacey Jennifer Garner Robbie Amell,Ch..

Understand Basic Information About the Data

In [ ]:
```
# Pandas provide many functions to understand the shape, number of column
# info( ) is one of my favorite methods that gives all necessary informat
```

info( )

In [19]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Rank                1000 non-null   int64  
 1   Title               1000 non-null   object 
 2   Genre               1000 non-null   object 
 3   Description         1000 non-null   object 
 4   Director            1000 non-null   object 
 5   Actors              1000 non-null   object 
 6   Year                1000 non-null   int64  
 7   Runtime (Minutes)   1000 non-null   int64  
 8   Rating              1000 non-null   float64
 9   Votes               1000 non-null   int64  
 10  Revenue (Millions)  872 non-null    float64
 11  Metascore           936 non-null    float64
dtypes: float64(3), int64(4), object(5)
memory usage: 93.9+ KB
```

shape

In [22]:
```
# shape can be used to get the shape of dataframe
data.shape
```

Out[22]:
```
(1000, 12)
```

columns

In [25]:
```
#columns gives us the list of columns in the dataframe
# This function tells us that there are 1000 rows and 12 columns in the d
data.columns
```

Out[25]:
```
Index(['Rank', 'Title', 'Genre', 'Description', 'Director', 'Actors', 'Y
ear',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore'],
      dtype='object')
```

describe( )

In [30]:
```
# describe( ) method gives the basic statistical summaries of all numeric

data.describe()
```

Out[30]:

	Rank	Year	Runtime (Minutes)	Rating	Votes	Re (Mi
count	1000.000000	1000.000000	1000.000000	1000.000000	1.000000e+03	872.0
mean	500.500000	2012.783000	113.172000	6.723200	1.698083e+05	82.9
std	288.819436	3.205962	18.810908	0.945429	1.887626e+05	103.2
min	1.000000	2006.000000	66.000000	1.900000	6.100000e+01	0.0
25%	250.750000	2010.000000	100.000000	6.200000	3.630900e+04	13.2
50%	500.500000	2014.000000	111.000000	6.800000	1.107990e+05	47.9
75%	750.250000	2016.000000	123.000000	7.400000	2.399098e+05	113.7
max	1000.000000	2016.000000	191.000000	9.000000	1.791916e+06	936.6

In [32]:
```
# Some Insights from the Description Table

# The min and max values in ‘Year’ depict the minimum and maximum release
# The average rating for the movies in this dataset is about 6.7 and the
# The maximum revenue earned by a movie is 936.6 million
```

Data Selection – Indexing and Slicing

Extract Data Using Columns

In [36]:
```
# Let’s quickly extract ‘Genre’ data from the dataframe
# Extract data as series
data['Genre']
```

Out[36]:
```
0       Action,Adventure,Sci-Fi
1      Adventure,Mystery,Sci-Fi
2               Horror,Thriller
3       Animation,Comedy,Family
4      Action,Adventure,Fantasy
                 ...
995         Crime,Drama,Mystery
996                      Horror
997         Drama,Music,Romance
998            Adventure,Comedy
999       Comedy,Family,Fantasy
Name: Genre, Length: 1000, dtype: object
```

In [38]:
```
# Extract data as dataframe
data[['Genre']]
```

Out[38]:

	Genre
0	Action,Adventure,Sci-Fi
1	Adventure,Mystery,Sci-Fi
2	Horror,Thriller
3	Animation,Comedy,Family
4	Action,Adventure,Fantasy
...	...
995	Crime,Drama,Mystery
996	Horror
997	Drama,Music,Romance
998	Adventure,Comedy
999	Comedy,Family,Fantasy

1000 rows × 1 columns

In [40]:
```
# If we want to extract multiple columns from the data, simply add the co
data[['Title','Genre','Actors','Director','Rating']]
```

Out[40]:

	Title	Genre	Actors	Director	Rating
0	Guardians of the Galaxy	Action,Adventure,Sci-Fi	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	James Gunn	8.1
1	Prometheus	Adventure,Mystery,Sci-Fi	Noomi Rapace, Logan Marshall-Green, Michael Fa...	Ridley Scott	7.0
2	Split	Horror,Thriller	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	M. Night Shyamalan	7.3
3	Sing	Animation,Comedy,Family	Matthew McConaughey,Reese Witherspoon, Seth Ma...	Christophe Lourdelet	7.2
4	Suicide Squad	Action,Adventure,Fantasy	Will Smith, Jared Leto, Margot Robbie, Viola D...	David Ayer	6.2
...	...	...	...	...	...
995	Secret in Their Eyes	Crime,Drama,Mystery	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	Billy Ray	6.2
996	Hostel: Part II	Horror	Lauren German, Heather Matarazzo, Bijou Philli...	Eli Roth	5.5
997	Step Up 2: The Streets	Drama,Music,Romance	Robert Hoffman, Briana Evigan, Cassie Ventura,...	Jon M. Chu	6.2
998	Search Party	Adventure,Comedy	Adam Pally, T.J. Miller, Thomas Middleditch,Sh...	Scot Armstrong	5.6
999	Nine Lives	Comedy,Family,Fantasy	Kevin Spacey, Jennifer Garner, Robbie Amell,Ch...	Barry Sonnenfeld	5.3

1000 rows × 5 columns

Extract Data Using Rows

In [43]:
```
# loc and iloc are two functions that can be used to slice data from spec

'''loc – locates the rows by name

loc performs slicing based explicit index.
```

```
It takes string indexes to retrieve data from specified rows
iloc – locates the rows by integer index

iloc performs slicing based on Python's default numerical index.'''
```

Out[43]:
```
'loc – locates the rows by name\n\nloc performs slicing based explicit i
ndex.\nIt takes string indexes to retrieve data from specified rows\nilo
c – locates the rows by integer index\n\niloc performs slicing based on
Python's default numerical index.'
```

In [53]:
```
# Read data with specified explicit index.
# We will use this later in our analysis
data_indexed = pd.read_csv('/Users/babarhussain/Documents/movies.csv', in
```

In [63]:
```
data_indexed.loc[['Suicide Squad','Split']][['Genre','Actors','Director',
```

Out[63]:

Title	Genre	Actors	Director	Rating	Revenue (Millions)
Suicide Squad	Action,Adventure,Fantasy	Will Smith, Jared Leto, Margot Robbie, Viola D...	David Ayer	6.2	325.02
Split	Horror,Thriller	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	M. Night Shyamalan	7.3	138.12

iloc

In [ ]:
```

```

In [66]:
```
data.iloc[10:15][['Title','Rating','Revenue (Millions)']]
```

Out[66]:

	Title	Rating	Revenue (Millions)
10	Fantastic Beasts and Where to Find Them	7.5	234.02
11	Hidden Figures	7.8	169.27
12	Rogue One	7.9	532.17
13	Moana	7.7	248.75
14	Colossal	6.4	2.87

In [ ]:
```

```

In [ ]:
```

```