This notebook is in line with the tutorial in https://www.learndatasci.com/tutorials/python-pandas-tutorial-complete-introduction-for-beginners/

```
from google.colab import drive
drive.mount('/content/gdrive')
```

```
Mounted at /content/gdrive
```

```
import pandas as pd
```

```
data = {
    'apples' : [0, 2, 1, 3],
    'oranges' : [1, 5, 2, 4]
}

purchases = pd.DataFrame(data);
purchases
```

	apples	oranges
0	0	1
1	2	5
2	1	2
3	3	4

```
movies_df = pd.read_csv("/content/gdrive/MyDrive/Suraksha/IMDB-Movie-Data.csv", index_col=
```

```
movies_df.head(5)
```

Title	Rank	Genre	Description	Director	Acto
Guardians of the Galaxy	1	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, V Diesel, Bradl Cooper, Zoe S
Prometheus	2	Adventure,Mystery,Sci-Fi	Following clues to the origin of	Ridley	Noomi Rapac Logan Marsha

```
movies_df.tail(5)
```

Title	Rank	Genre	Description	Director	Actors	Year	R
Secret in Their Eyes	996	Crime,Drama,Mystery	A tight-knit team of rising investigators, alo...	Billy Ray	Chiwetel Ejiofor, Nicole Kidman, Julia Roberts...	2015	
Hostel: Part II	997	Horror	Three American college students studying abroa...	Eli Roth	Lauren German, Heather Matarazzo, Bijou Philli...	2007	
Step Up 2: The	998	Drama,Music,Romance	Romantic sparks occur between two dance	Jon M. Chu	Robert Hoffman, Briana Evigan, Cassie	2008	

```
movies_df.shape
```

```
(1000, 11)
```

```
movies_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1000 entries, Guardians of the Galaxy to Nine Lives
Data columns (total 11 columns):
 #   Column              Non-Null Count  Dtype  
---  ------              --------------  -----  
 0   Rank                1000 non-null   int64  
 1   Genre               1000 non-null   object 
 2   Description         1000 non-null   object 
 3   Director            1000 non-null   object 
 4   Actors              1000 non-null   object 
 5   Year                1000 non-null   int64  
 6   Runtime (Minutes)   1000 non-null   int64  
 7   Rating              1000 non-null   float64
 8   Votes               1000 non-null   int64  
 9   Revenue (Millions)  872 non-null    float64
```

```
 10  Metascore           936 non-null    float64
dtypes: float64(3), int64(4), object(4)
memory usage: 93.8+ KB
```

```
movies_df = movies_df.drop_duplicates(keep = 'first')
```

```
movies_df.columns
```

```
Index(['Rank', 'Genre', 'Description', 'Director', 'Actors', 'Year',
       'Runtime (Minutes)', 'Rating', 'Votes', 'Revenue (Millions)',
       'Metascore'],
      dtype='object')
```

```
movies_df.rename(columns = {'Runtime (Minutes)' : 'Runtime', 'Revenue (Millions)' : 'Reven
movies_df.columns
```

```
Index(['Rank', 'Genre', 'Description', 'Director', 'Actors', 'Year', 'Runtime',
       'Rating', 'Votes', 'Revenue_millions', 'Metascore'],
      dtype='object')
```

```
movies_df.isnull().sum()
```

```
Rank                  0
Genre                 0
Description           0
Director              0
Actors                0
Year                  0
Runtime               0
Rating                0
Votes                 0
Revenue_millions    128
Metascore            64
dtype: int64
```

```
movies_dfTmp = movies_df.dropna(axis=0)
movies_dfTmp.shape
```

```
(838, 11)
```

```
movies_dfTmp = movies_df.dropna(axis=1) #To drop columns containing null values
movies_dfTmp.shape
```

```
(1000, 9)
```

```
movies_df.shape
```

```
(1000, 11)
```

```
revenue = movies_df['Revenue_millions']
revenue.head(5)
```

```
Title
Guardians of the Galaxy    333.13
Prometheus                 126.46
Split                      138.12
Sing                       270.32
Suicide Squad              325.02
Name: Revenue_millions, dtype: float64
```

```
meanRev = revenue.mean(0)
revenue.fillna(meanRev, inplace=True)
movies_df.isnull().sum()
```

```
Rank                 0
Genre                0
Description          0
Director             0
Actors               0
Year                 0
Runtime              0
Rating               0
Votes                0
Revenue_millions     0
Metascore           64
dtype: int64
```

```
movies_df.describe()
```

	Rank	Year	Runtime	Rating	Votes	Revenue_mi
count	1000.000000	1000.000000	1000.000000	1000.000000	1.000000e+03	1000.(
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	96.4
min	1.000000	2006.000000	66.000000	1.900000	6.100000e+01	0.(
25%	250.750000	2010.000000	100.000000	6.200000	3.630900e+04	17.4
50%	500.500000	2014.000000	111.000000	6.800000	1.107990e+05	60.3
75%	750.250000	2016.000000	123.000000	7.400000	2.399098e+05	99.1
max	1000.000000	2016.000000	191.000000	9.000000	1.791916e+06	936.6

```
movies_df['Genre'].value_counts()
```

```
Action,Adventure,Sci-Fi    50
Drama                      48
Comedy,Drama,Romance       35
Comedy                     32
Drama,Romance              31
                           ..
Comedy,Horror,Romance       1
Drama,History,War           1
Comedy,Sci-Fi               1
Crime,Thriller              1
Animation,Drama,Romance     1
Name: Genre, Length: 207, dtype: int64
```

```
movies_df.corr()
```

	Rank	Year	Runtime	Rating	Votes	Revenue_milli
Rank	1.000000	-0.261605	-0.221739	-0.219555	-0.283876	-0.252
Year	-0.261605	1.000000	-0.164900	-0.211219	-0.411904	-0.117
Runtime	-0.221739	-0.164900	1.000000	0.392214	0.407062	0.247
Rating	-0.219555	-0.211219	0.392214	1.000000	0.511537	0.189
Votes	-0.283876	-0.411904	0.407062	0.511537	1.000000	0.607
Revenue_millions	-0.252996	-0.117562	0.247834	0.189527	0.607941	1.000
Metascore	-0.191869	-0.079305	0.211978	0.631897	0.325684	0.133

```
subset = movies_df[['Genre', 'Rating']]
type(subset)
```

```
pandas.core.frame.DataFrame
```

```
movies_df.loc['Prometheus']
movies_df.iloc[1]
```

```
Rank                                                        2
Genre                                    Adventure,Mystery,Sci-Fi
Description         Following clues to the origin of mankind, a te...
Director                                           Ridley Scott
Actors              Noomi Rapace, Logan Marshall-Green, Michael Fa...
Year                                                     2012
Runtime                                                   124
Rating                                                      7
Votes                                                  485820
Revenue_millions                                       126.46
Metascore                                                  65
Name: Prometheus, dtype: object
```

```
movie_subset = movies_df.iloc[1:4]
movie_subset
```

Title	Rank	Genre	Description	Director	Acto
Prometheus	2	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapac Logan Marsha Green, Michael Fa
Split	3	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, An Taylor-Joy, Haley L Richar

```
rating = movies_df['Rating']
rating[rating.gt(8.5)]
```

```
Title
Interstellar        8.6
The Dark Knight     9.0
Inception           8.8
Kimi no na wa       8.6
Dangal              8.8
The Intouchables    8.6
Name: Rating, dtype: float64
```

```
moviesByRidley = movies_df[(movies_df['Director'] == "Ridley Scott") & movies_df['Rating']
moviesByRidley.head(4)
```

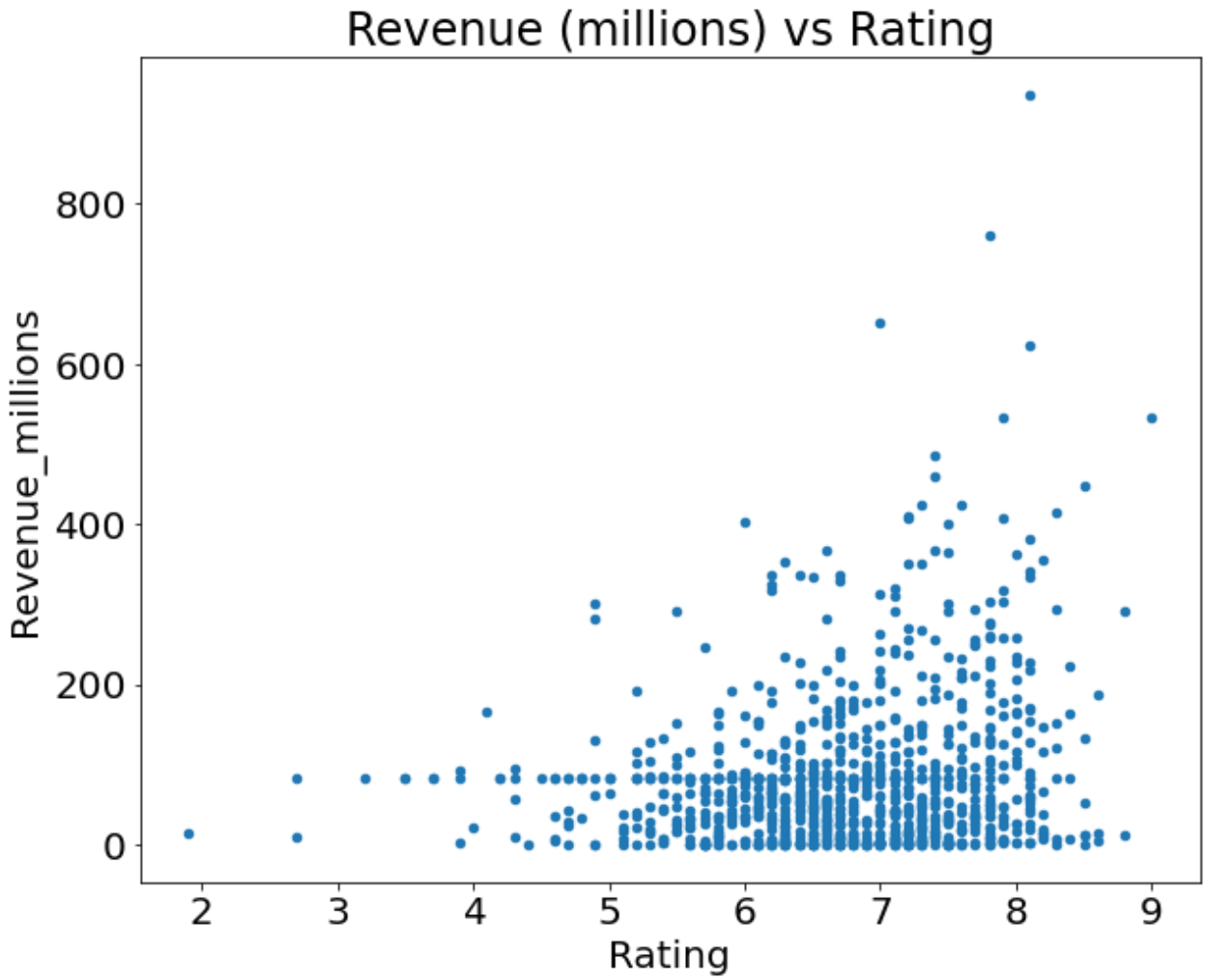
Title	Rank	Genre	Description	Director	Actors	Year	Run
The Martian	103	Adventure,Drama,Sci-Fi	An astronaut becomes stranded on Mars after hi...	Ridley Scott	Matt Damon, Jessica Chastain, Kristen Wiig, Ka...	2015	

```
movies_df[
    ((movies_df['Year'] >= 2005) & (movies_df['Year'] <= 2010))
    & (movies_df['Rating'] > 8.0)
    & (movies_df['Revenue_millions'] < movies_df['Revenue_millions'].quantile(0.25))
]
```

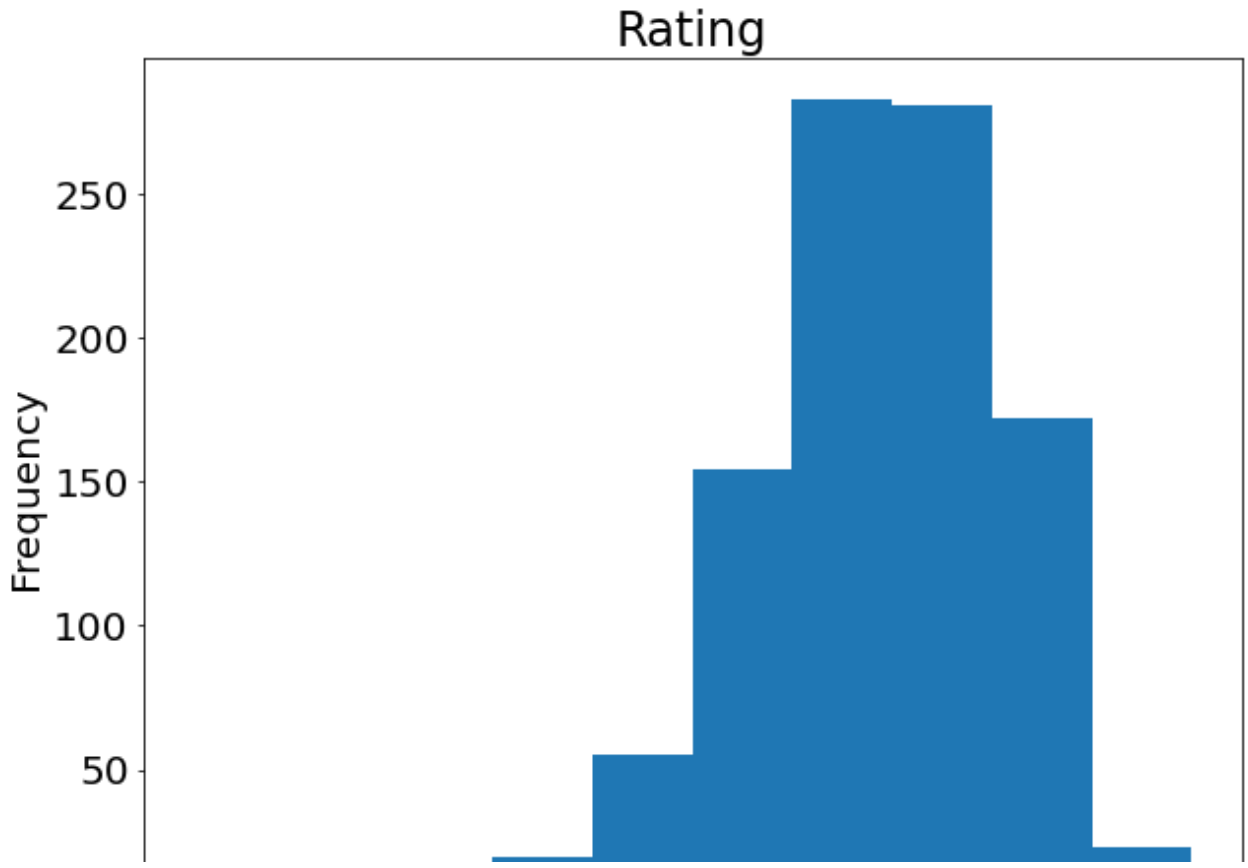
Title	Rank	Genre	Description	Director	Actors	Yea
3 Idiots	431	Comedy,Drama	Two friends are searching for their long lost ...	Rajkumar Hirani	Aamir Khan, Madhavan, Mona Singh, Sharman Joshi	200
The Lives of Others	477	Drama,Thriller	In 1984 East Berlin, an agent of the secret po...	Florian Henckel von Donnersmarck	Ulrich Mühe, Martina Gedeck,Sebastian Koch, Ul...	200

```
import matplotlib.pyplot as plt
plt.rcParams.update({'font.size': 20, 'figure.figsize': (10, 8)})
```

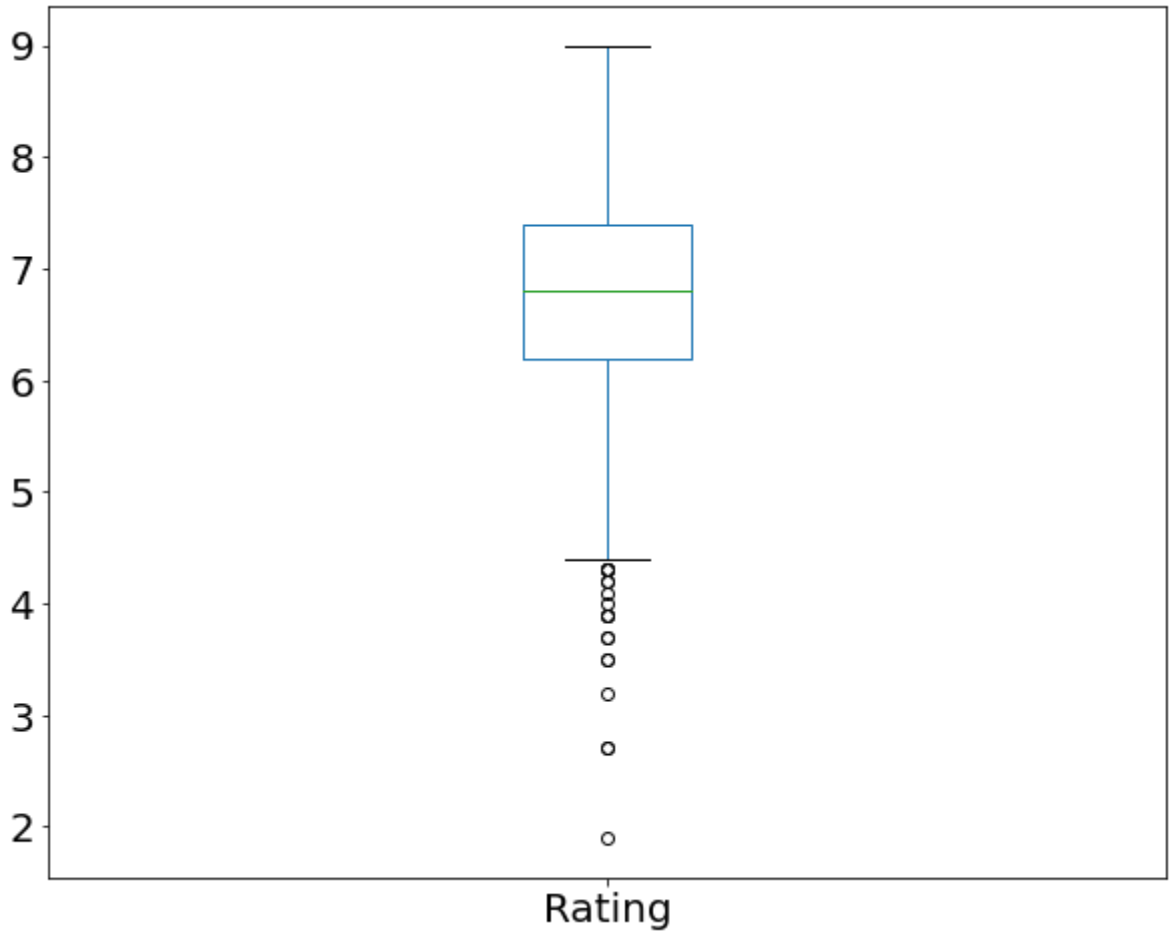
```
movies_df.plot(kind='scatter', x='Rating', y='Revenue_millions', title='Revenue (millions)
```




```
movies_df['Rating'].plot(kind='hist', title='Rating');
```




```
movies_df['Rating'].plot(kind="box");
```



✓ 0s completed at 5:17 PM