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

```
    Drive already mounted at /content/gdrive; to attempt to forcibly remount, call drive.mount("/content/gdrive", force_remount=True).
```

```
import pandas as pd
```

```
DatasetBaseFolder = '/content/gdrive/MyDrive/Colab Notebooks/AIpandas/'
```

```
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(DatasetBaseFolder+"IMDB-Movie-Data.csv", index_col="Title")
```

```
#lets see first 5 rows
movies_df.head(5)
```

	Rank	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
Title											
Guardians of the Galaxy	1	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
Prometheus	2	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
Split	3	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
Sing	4	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
Suicide Squad	5	Action,Adventure,Fantasy	A secret government agency	David Ayer	Will Smith, Jared Leto, Margot Robbie,	2016	123	6.2	393727	325.02	40.0

```
#Lets see last 5 rows
movies_df.tail(5)
```

	Rank	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascor
Title											
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	111	6.2	27585	NaN	45
Hostel: Part II	997	Horror	Three American college students	Eli Roth	Lauren German, Heather Matarazzo	2007	94	5.5	73152	17.54	46

```
movies_df.shape
```

```
    (1000, 11)
```

```
#To get an overview of the dataset
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
```

```
#If you want to remove duplicate instances
movies_df = movies_df.drop_duplicates(keep = 'first') #Drop all instances keep = false inplace=True
```

```
#If you wish to rename columns
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)' : 'Revenue_millions'}, inplace=True)
movies_df.columns
```

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

```
#To count number of null entries in each colum
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) #To drop instances with null values
movies_dfTmp.shape
#movies_df.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)
```

```
#Imputing with Mean
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() #Note that this get updated
```

```
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
```

```
#Describ the Dataset
movies_df.describe()
```

	Rank	Year	Runtime	Rating	Votes	Revenue_millions	Metascore
count	1000.000000	1000.000000	1000.000000	1000.000000	1.000000e+03	1000.000000	936.000000
mean	500.500000	2012.783000	113.172000	6.723200	1.698083e+05	82.956376	58.985043
std	288.819436	3.205962	18.810908	0.945429	1.887626e+05	96.412043	17.194757
min	1.000000	2006.000000	66.000000	1.900000	6.100000e+01	0.000000	11.000000
25%	250.750000	2010.000000	100.000000	6.200000	3.630900e+04	17.442500	47.000000
50%	500.500000	2014.000000	111.000000	6.800000	1.107990e+05	60.375000	59.500000
75%	750.250000	2016.000000	123.000000	7.400000	2.399098e+05	99.177500	72.000000
max	1000.000000	2016.000000	191.000000	9.000000	1.791916e+06	936.630000	100.000000

```
#if you want to count
movies_df['Genre'].value_counts()
```

```
Action,Adventure,Sci-Fi     50
Drama                       48
Comedy,Drama,Romance        35
Comedy                      32
Drama,Romance               31
                            ..
Drama,Family,Fantasy         1
Action,Comedy,Mystery        1
Comedy,Western               1
Mystery,Romance,Thriller     1
Comedy,Romance,Western       1
Name: Genre, Length: 207, dtype: int64
```

```
#Correlation
movies_df.corr()  #Note the attributes in  S
```

	Rank	Year	Runtime	Rating	Votes	Revenue_millions	Metascore
Rank	1.000000	-0.261605	-0.221739	-0.219555	-0.283876	-0.252996	-0.191869
Year	-0.261605	1.000000	-0.164900	-0.211219	-0.411904	-0.117562	-0.079305
Runtime	-0.221739	-0.164900	1.000000	0.392214	0.407062	0.247834	0.211978
Rating	-0.219555	-0.211219	0.392214	1.000000	0.511537	0.189527	0.631897
Votes	-0.283876	-0.411904	0.407062	0.511537	1.000000	0.607941	0.325684
Revenue_millions	-0.252996	-0.117562	0.247834	0.189527	0.607941	1.000000	0.133328
Metascore	-0.191869	-0.079305	0.211978	0.631897	0.325684	0.133328	1.000000

```
#slicing along columns
subset = movies_df[['Genre', 'Rating']]
type(subset)
```

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

```
#Slicing along rows
movies_df.loc['Prometheus'] #using key index
movies_df.iloc[1] #using numerical index
```

```
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
```

```
#few instances 1 through 3
movie_subset = movies_df.iloc[1:4]
movie_subset
```

	Rank	Genre	Description	Director	Actors	Year	Runtime	Rating	Votes	Revenue_mill
Title										
Prometheus	2	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	1

```
#conditional selection
#Pick movies with rating more than 8.5
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
```

```
#Pick movies based on Director
moviesByRidley = movies_df[(movies_df['Director'] == "Ridley Scott") & movies_df['Rating'].gt(7.5)]
moviesByRidley.head(4)
```

	Rank	Genre	Description	Director	Actors	Year	Runtime	Rating	Votes	Revenue_millions	Metas
Title											
The Martian	103	Adventure,Drama,Sci-Fi	An astronaut becomes stranded on	Ridley Scott	Matt Damon, Jessica Chastain,	2015	144	8.0	556097	228.43	

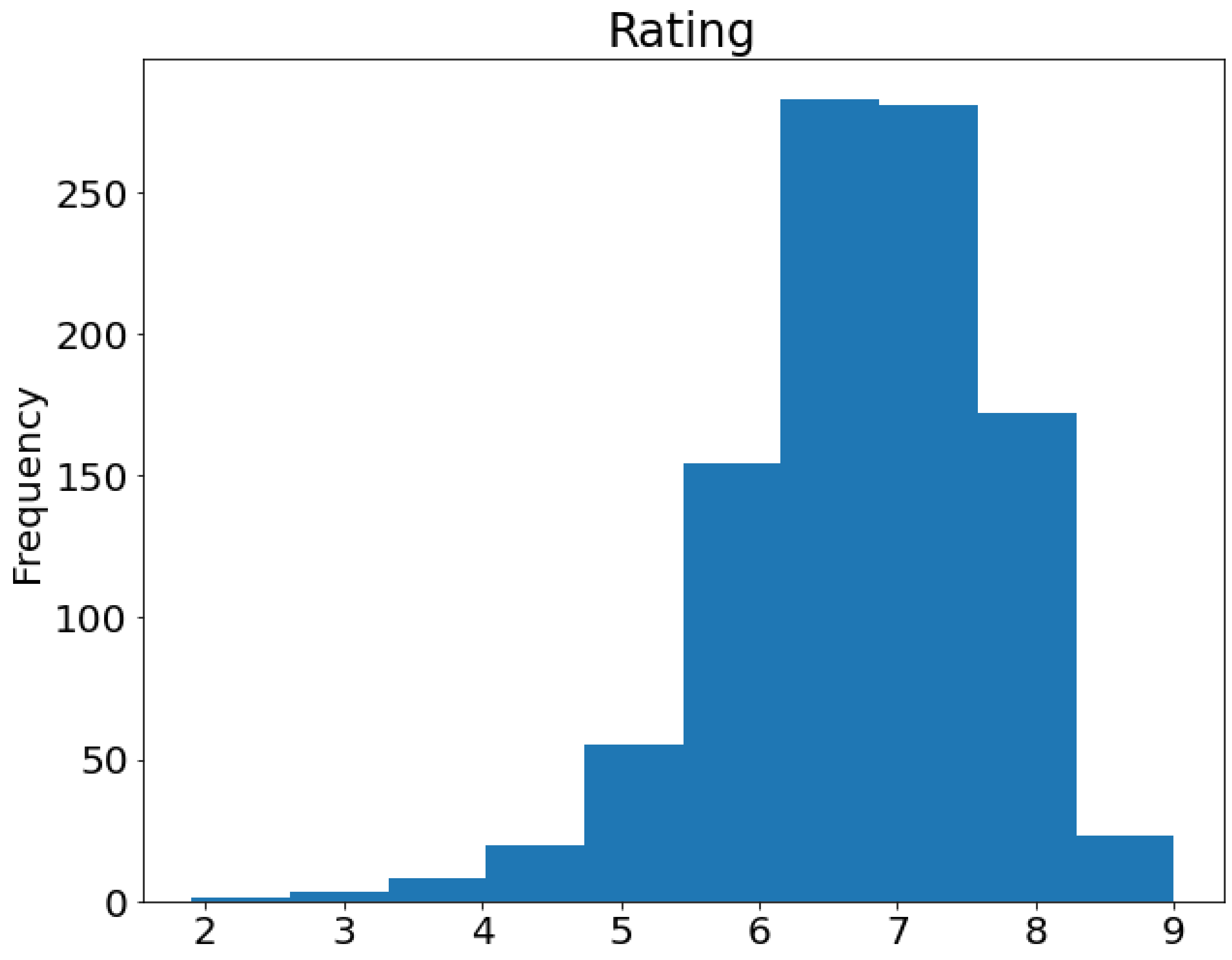
```
#all movies that were released between 2005 and 2010, have a rating above 8.0, but made below the 25th percentile in revenue.
movies_df[
    ((movies_df['Year'] >= 2005) & (movies_df['Year'] <= 2010))
    & (movies_df['Rating'] > 8.0)
    & (movies_df['Revenue_millions'] < movies_df['Revenue_millions'].quantile(0.25))
]
```

	Rank	Genre	Description	Director	Actors	Year	Runtime	Rating	Votes	Revenue_millions
Title										
3 Idiots	431	Comedy,Drama	Two friends are searching for their long lost ...	Rajkumar Hirani	Aamir Khan, Madhavan, Mona Singh, Sharman Joshi	2009	170	8.4	238789	6.52
The Lives of Others	477	Drama,Thriller	In 1984 East Berlin, an agent of the	Florian Henckel von	Ulrich Mühe, Martina Gedeck Sebastian	2006	137	8.5	278103	11.28

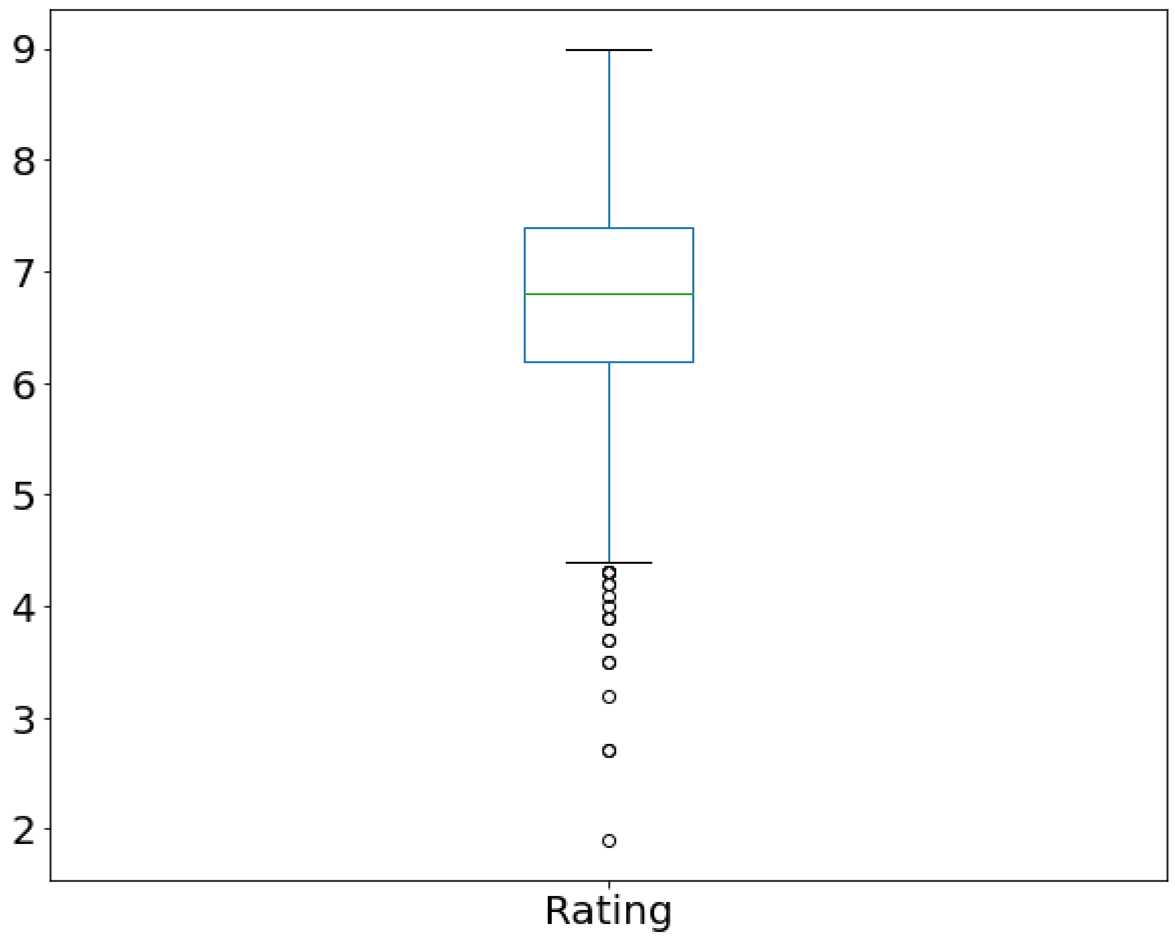
```
import matplotlib.pyplot as plt
plt.rcParams.update({'font.size': 20, 'figure.figsize': (10, 8)})
```

```
#For categorical variables utilize Bar Charts* and Boxplots.
#For continuous variables utilize Histograms, Scatterplots, Line graphs, and Boxplots.
movies_df.plot(kind='scatter', x='Rating', y='Revenue_millions', title='Revenue (millions) vs Rating');
```

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



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



✓ 0s completed at 16:51