```
## DMW A1: Data Preprocessing
## Aditya Agre TYCOA6

## Data set : movies.csv


import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


df = pd.read_csv('movies2.csv')
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	NaN

5) Going through all measures of central tendency

Studying modal values per column
```
df.mode(axis = 0)
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime
0	Bleach: Burîchi	(2020–)	\nComedy	7.2	\nAdd a Plot\n	\n	7	24.0

Studying mean values per column
```
df.mean()
```

```
/var/folders/ft/m0h88bl55gl0qmgjxz9qczfc0000gn/T/ipykernel_4020/2762134590.
  df.mean()
RATING      6.921176
RunTime    68.688539
dtype: float64
```

Studying median values per column
```
df.median()
```

```
/var/folders/ft/m0h88bl55gl0qmgjxz9qczfc0000gn/T/ipykernel_4020/1181268256.
  df.median()
RATING      7.1
RunTime    60.0
dtype: float64
```

Handling Missing Values

which of these columns have null values

```
no_of_col = df.shape[1]

for i in range(no_of_col):
    if(df[:i].isnull().values.any()):
        print("Col ",i," has null values.")
```

```
Col  5  has null values.
Col  6  has null values.
Col  7  has null values.
```

```
## Checking number of null entries per column.
df.isnull().sum()
```

```
MOVIES         0
YEAR         644
GENRE         80
RATING      1820
ONE-LINE       0
STARS          0
VOTES       1820
RunTime     2958
dtype: int64
```

```
# RunTime has null values.
# It is also a numerical column
# So we can replace the null values with the median

df['RunTime'] = df['RunTime'].fillna(df['RunTime'].median())
df
## As you can see the RunTime column now doesnt show any null values like before
## All the null values have been replaced with mode value 24
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	NaN	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	NaN

```
# Similarly:
df['RATING'].replace(np.NaN, df['RATING'].mode()[0],inplace=True)
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	7.2	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	NaN

```
# Similarly:
df['YEAR'].replace(np.NaN, df['YEAR'].mode()[0],inplace=True)
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	7.2	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	NaN

```
# Similarly:
df['VOTES'].replace(np.NaN, df['VOTES'].mode()[0],inplace=True)
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	7.2	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	7

```
df.isnull().sum()
```

```
MOVIES       0
YEAR         0
GENRE       80
RATING       0
ONE-LINE     0
STARS        0
VOTES        0
RunTime      0
dtype: int64
```

```
# Genre is a non-numerical datatype
# 80 entries have null genre

# Removing these entries

df.dropna(subset=['GENRE'], inplace=True)
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	7.2	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	7

```
df.isnull().sum()
```

```
MOVIES      0
YEAR        0
GENRE       0
RATING      0
ONE-LINE    0
STARS       0
VOTES       0
RunTime     0
dtype: int64
```

```
# No more NULL values
```

```
## Removal of duplicates
```

```
# Now lets handle the duplicates

df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010– 2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	7.2	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n	7

```
# We currently have 9919 rows
# Lets look at the duplicate rows now
df[df.duplicated(keep='first')]
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES	RunTime
6833	Mighty Little Bheem	(2019–)	\nAnimation, Short, Adventure	7.2	\nAdd a Plot\n	\n Directors:\nRajiv Chilaka, \nKrishna Moh...	7	60.0
6835	Mighty Little Bheem	(2019–)	\nAnimation, Short, Adventure	9.0	\nAdd a Plot\n	\n Directors:\nRajiv Chilaka, \nKrishna Moh...	6	60.0
6836	Mighty Little Bheem	(2019–)	\nAnimation, Short, Adventure	9.0	\nAdd a Plot\n	\n Directors:\nRajiv Chilaka, \nKrishna Moh...	6	60.0
6837	Mighty Little Bheem	(2019–)	\nAnimation, Short, Adventure	7.2	\nAdd a Plot\n	\n Directors:\nRajiv Chilaka, \nKrishna Moh...	7	60.0
	Mighty	(2019–	\nAnimation,		\nAdd	\n Directors:\nRajiv		

```
## 429 repeating rows

# Removing these rows
df.drop_duplicates(keep='first', inplace=True)
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
0	Blood Red Sky	-2021	\nAction, Horror, Thriller	6.1	\nA woman with a mysterious illness is forced ...	\n Director:\nPeter Thorwarth\nl \n Star...	21,062
1	Masters of the Universe: Revelation	(2021–)	\nAnimation, Action, Adventure	5.0	\nThe war for Eternia begins again in what may...	\n \n Stars:\nChris Wood, \nSara...	17,870
2	The Walking Dead	(2010–2022)	\nDrama, Horror, Thriller	8.2	\nSheriff Deputy Rick Grimes wakes up from a c...	\n \n Stars:\nAndrew Lincoln, \n...	8,85,805
3	Rick and Morty	(2013–)	\nAnimation, Adventure, Comedy	9.2	\nAn animated series that follows the exploits...	\n \n Stars:\nJustin Roiland, \n...	4,14,849
4	Army of Thieves	-2021	\nAction, Crime, Horror	7.2	\nA prequel, set before the events of Army of	\n Director:\nMatthias Schweighöfer\nl \n ...	7

```
df.shape[0]
```

```
9490
```

```
# 9490 rows left
```

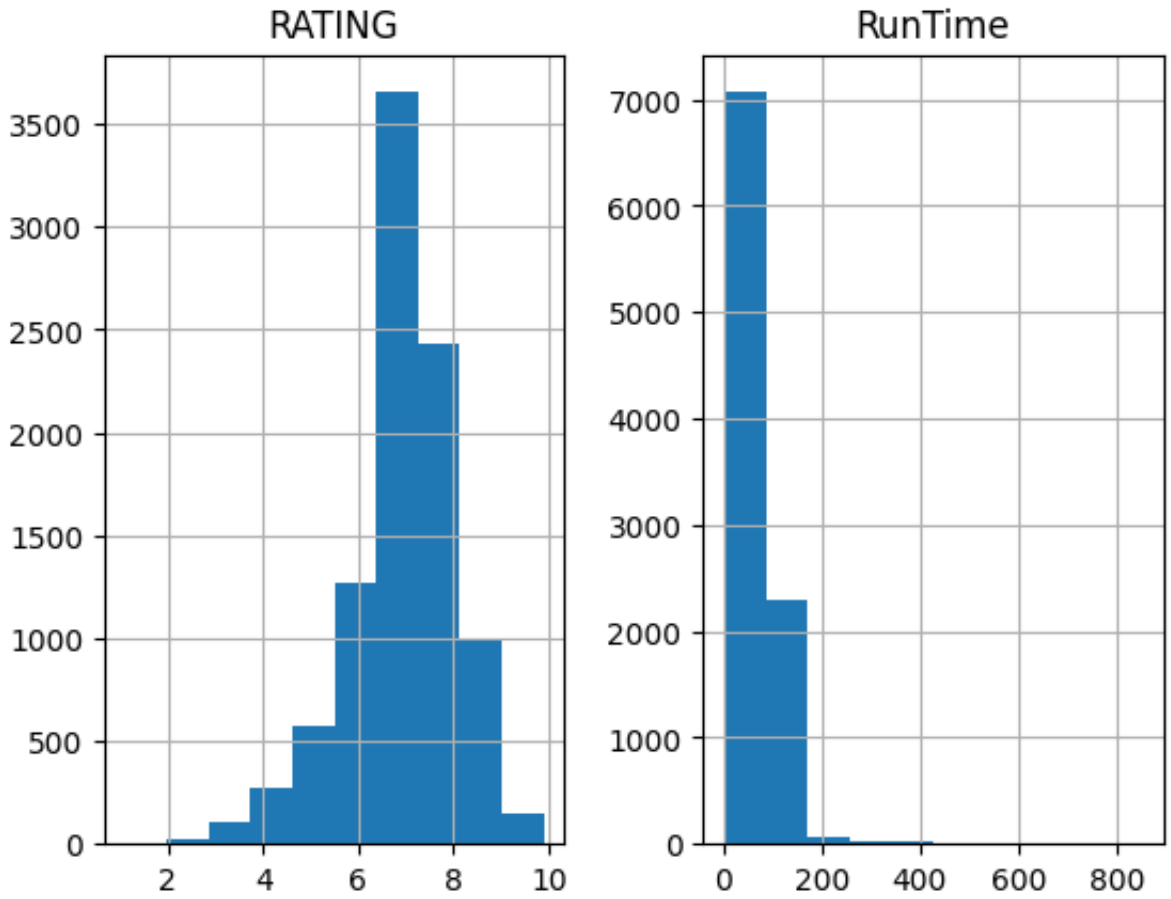
```
## data has not been arranged according to any pattern
## Yet, we are shuffling

df = df.sample(frac=1, random_state=42)
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
2211	The Numbers Station	-2013	\nAction, Thriller	5.6	\nA disgraced black ops agent is dispatched to...	\n Director:\nKasper Barfoed\nl \n Stars...	19,590
2908	The Stranded	(2019–)	\nAdventure, Drama, Mystery	6.3	\nWhen a tsunami strands dozens of teens on an...	\n \n Stars:\nPapangkorn Lerkcha...	1,081
1530	Spaceman of Bohemia	(2020–)	\nAdventure, Drama, Sci-Fi	7.2	\nJakub Procházka, who orphaned as a boy and r...	\n Director:\nJohan Renck\nl \n Stars:\n...	7
6542	Avatar: The Last Airbender	(2005–2008)	\nAnimation, Action, Adventure	9.1	Aang relives the events after finding out that...	\n Director:\nLauren MacMullan\nl \n Sta...	3,555
4080	Les gars sûrs	-2022	\nAction, Comedy	7.2	Plot Unknown, Sequel of De l'autre côté du pér	\n Director:\nLouis Leterrier\nl \n Star...	7

```
## Normalising Data using min max scaling
```

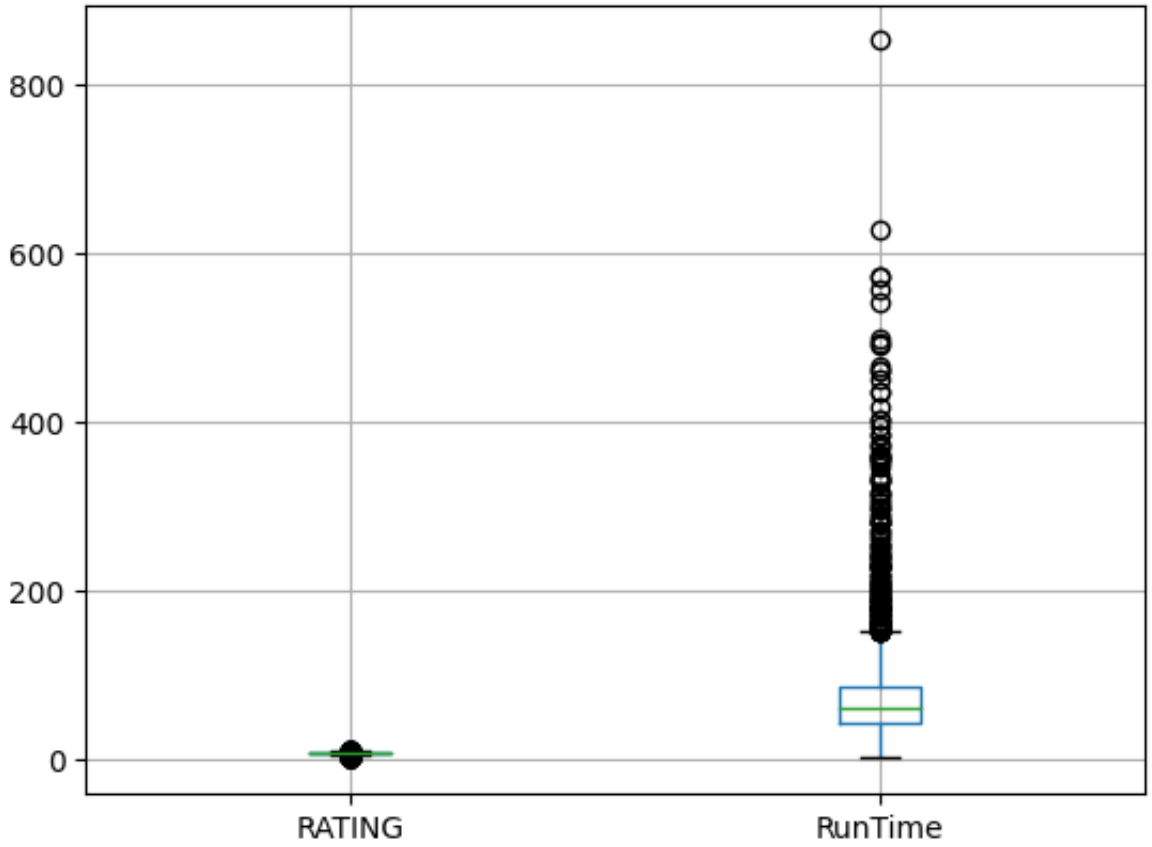
```
## PLotting histogram

df.hist()
```

```
array([[<AxesSubplot:title={'center':'RATING'}>,
        <AxesSubplot:title={'center':'RunTime'}>]], dtype=object)
```



```
## Plotting boxplot

df.boxplot()
```

```
<AxesSubplot:>
```




```
# Boxplot shows many outliers


# lets make a general func that removes outliers for a given column
```

```
def outlier_remove(col_i):
    ## Lets follow inter quartile range method
    ## return the values of the lower range limit and upper range limit
    ## these limits have values 1.5*(inter quartile range beyond) first and thir

    ## To find quartiles, we must sort the column
    sorted(col_i)

    Q1,Q3 = np.percentile(col_i , [25,75])
    ## because we are taking quartiles. therefore 25% and 75%
    inter_q_range = Q3-Q1

    l_lim = Q1 - (1.5 * inter_q_range)
    up_lim = Q3 + (1.5 * inter_q_range)
    return l_lim,up_lim, Q1,Q3, inter_q_range
```

```
l,u,q1,q3,iqr=outlier_remove(df.RunTime)
print(l,u,q1,q3,iqr)

rows = df.shape[0]

df.drop(df[(df.RunTime < l) | (df.RunTime > u)].index,inplace=True)
```

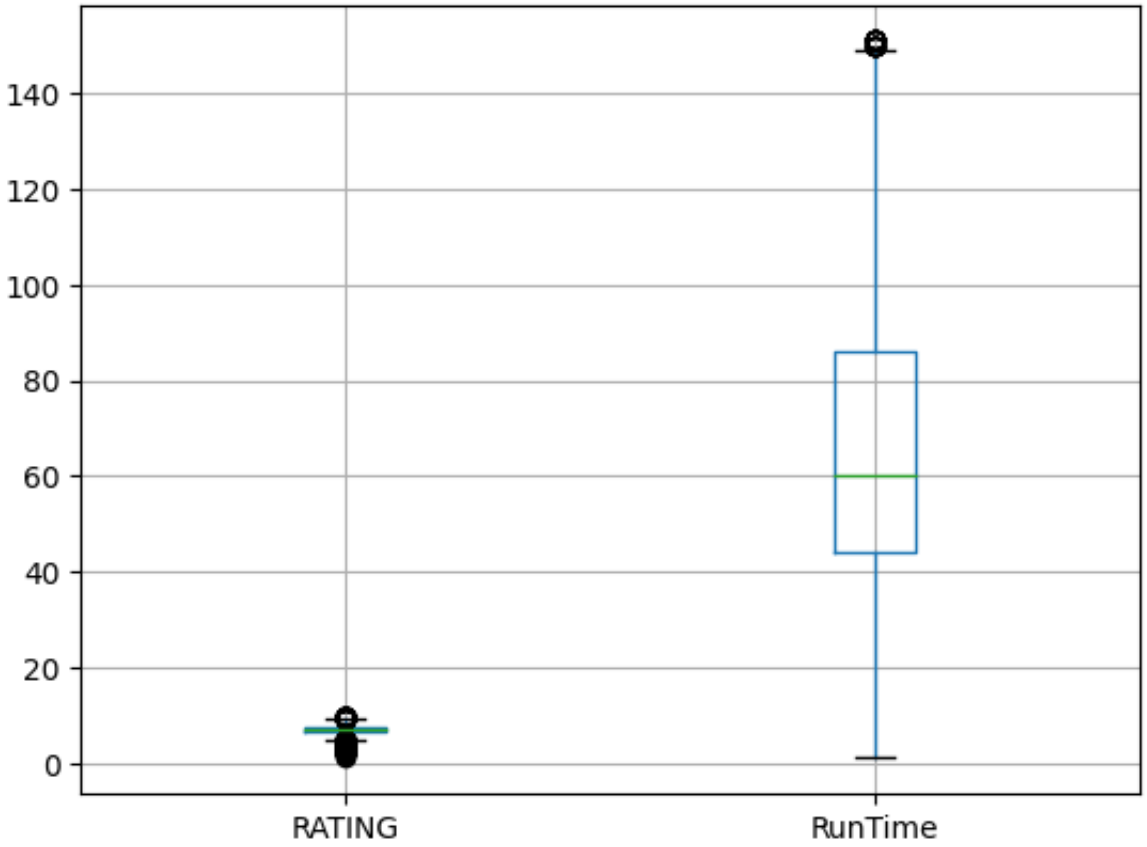
```
-20.5 151.5 44.0 87.0 43.0
```

```
df
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
2211	The Numbers Station	-2013	\nAction, Thriller	5.6	\nA disgraced black ops agent is dispatched to...	\n Director:\nKasper Barfoed\nl \n Stars...	19,590
2908	The Stranded	(2019–)	\nAdventure, Drama, Mystery	6.3	\nWhen a tsunami strands dozens of teens on an...	\n \n Stars:\nPapangkorn Lerkcha...	1,081
1530	Spaceman of Bohemia	(2020–)	\nAdventure, Drama, Sci-Fi	7.2	\nJakub Procházka, who orphaned as a boy and r...	\n Director:\nJohan Renck\nl \n Stars:\n...	7
6542	Avatar: The Last Airbender	(2005–2008)	\nAnimation, Action, Adventure	9.1	Aang relives the events after finding out that...	\n Director:\nLauren MacMullan\nl \n Sta...	3,555
4080	Les gars sûrs	-2022	\nAction, Comedy	7.2	Plot Unknown, Sequel of De l'autre côté du pér	\n Director:\nLouis Leterrier\nl \n Star...	7

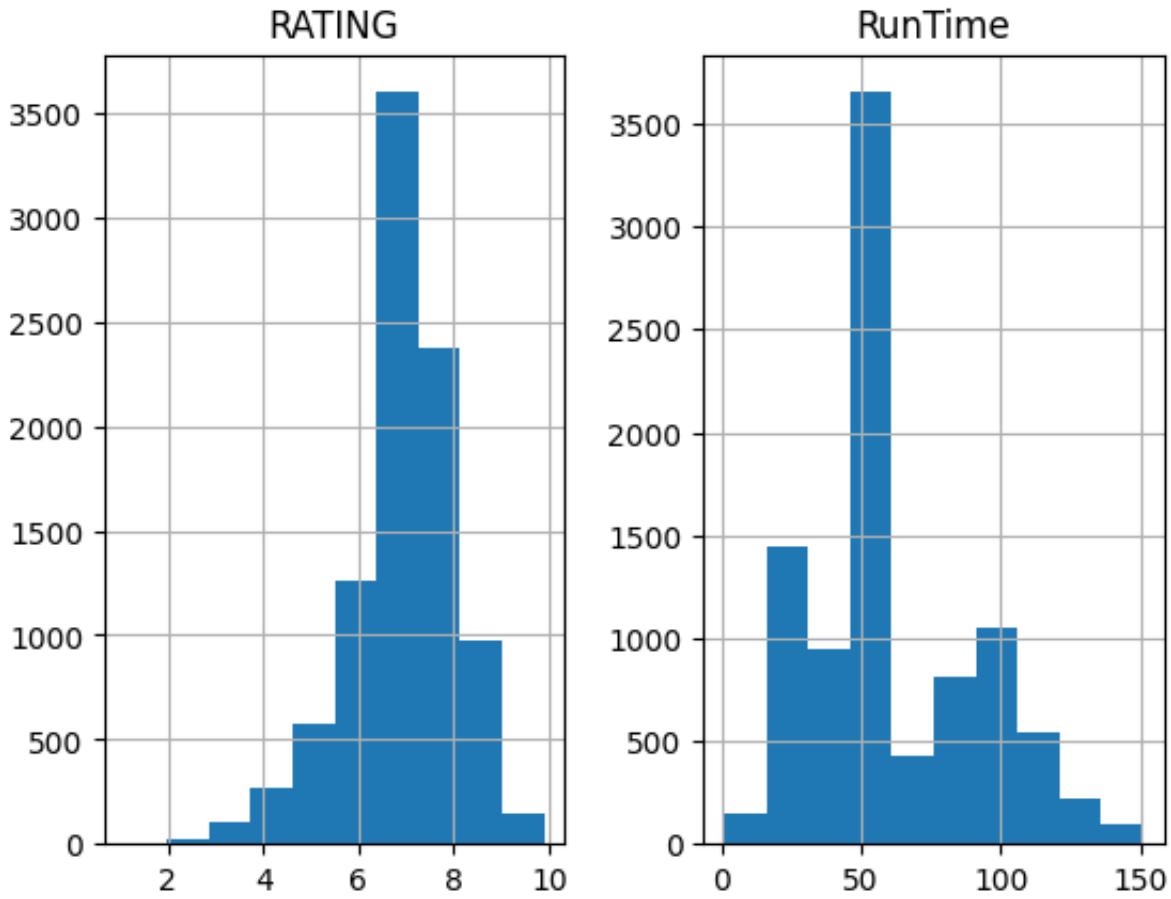
```
## Therefore outliers removed

df.boxplot()
```

```
<AxesSubplot:>
```



```
df.hist()
```

```
array([[<AxesSubplot:title={'center':'RATING'}>,
        <AxesSubplot:title={'center':'RunTime'}>]], dtype=object)
```



Double-click (or enter) to edit

```
# Scaling
```

```
from sklearn.preprocessing import MinMaxScaler

MMscaler = MinMaxScaler()
## We have created an object of minmax scaler class
```

```
df_ncol= df.select_dtypes(exclude=['object'])
df_ncol
## We are doing this to extract onlythe numeric cols
## A new dataframe is being created with just these cols
```

	RATING	RunTime
2211	5.6	89.0
2908	6.3	40.0
1530	7.2	60.0
6542	9.1	25.0
4080	7.2	60.0
...	...	...
5761	4.8	83.0
5204	6.4	70.0
5409	6.8	62.0
860	7.0	127.0
7380	7.0	42.0

9332 rows × 2 columns

```
type(df_ncol)
```

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

```
df_ncol.columns # to get column headers
```

```
Index(['RATING', 'RunTime'], dtype='object')
```

```
## Making a copy of df
temp = df

# Scalable columns
cols= df_ncol.columns

## Performing min max scalning
temp[cols]=  MMscaler.fit_transform(df[cols])
```

```
temp
```

	MOVIES	YEAR	GENRE	RATING	ONE-LINE	STARS	VOTES
2211	The Numbers Station	-2013	\nAction, Thriller	0.511364	\nA disgraced black ops agent is dispatched to...	\n Director:\nKasper Barfoed\nl \n Stars...	19,590
2908	The Stranded	(2019–)	\nAdventure, Drama, Mystery	0.590909	\nWhen a tsunami strands dozens of teens on an...	\n \n Stars:\nPapangkorn Lerkcha...	1,081
1530	Spaceman of Bohemia	(2020–)	\nAdventure, Drama, Sci-Fi	0.693182	\nJakub Procházka, who orphaned as a boy and r...	\n Director:\nJohan Renck\nl \n Stars:\n...	7
6542	Avatar: The Last Airbender	(2005–2008)	\nAnimation, Action, Adventure	0.909091	Aang relives the events after finding out that...	\n Director:\nLauren MacMullan\nl \n Sta...	3,555
4080	Les gars sûrs	-2022	\nAction, Comedy	0.693182	Plot Unknown, Sequel of De l'autre côté du pér	\n Director:\nLouis Leterrier\nl \n Star...	7

```
## RunTime and rating have now been chaged to values between 0 and 1


## Therefore, data has been preprocessed by handling missing values, duplicate v
```