12106692

August 31, 2023

1 K21UT CA1

- ***Registration Number : 12106692***
- ***Roll No : 09***
- ***Name : Garvit Joshi***
- ***Group : 1***

[20]:
```
import warnings

# Suppress the warning message
warnings.filterwarnings("ignore")
```

1.1 1.Find the null values from the data set and remove the null value if it is numeric and if it is character replace it with NAN

[21]:
```
import pandas as pd
import numpy as np

# Load your dataset into a pandas DataFrame (replace 'data.csv' with your file)
data = pd.read_csv('data.csv')

# Find null values
null_counts = data.isnull().sum()
data
```

[21]:
```
                                          Title  title_year     budget  \
0                                   La La Land        2016   30000000
1                                     Zootopia        2016  150000000
2                                         Lion        2016   12000000
3                                      Arrival        2016   47000000
4                       Manchester by the Sea        2016    9000000
..                                           …          …          …
95                                    Whiplash        2014    3300000
96                             Before Midnight        2013    3000000
97     Star Wars: Episode VII - The Force Awakens        2015  245000000
98  Harry Potter and the Deathly Hallows: Part I        2010  150000000
99                         Tucker and Dale vs Evil        2010    5000000
```

	Gross	actor_1_name	actor_2_name	\
0	151101803	Ryan Gosling	Emma Stone	
1	341268248	Ginnifer Goodwin	Jason Bateman	
2	51738905	Dev Patel	Nicole Kidman	
3	100546139	Amy Adams	Jeremy Renner	
4	47695371	Casey Affleck	Michelle Williams	
..	…	…	…	
95	13092000	J.K. Simmons	Melissa Benoist	
96	8114507	Seamus Davey-Fitzpatrick	Ariane Labed	
97	936662225	Doug Walker	Rob Walker	
98	296347721	Rupert Grint	Toby Jones	
99	223838	Katrina Bowden	Tyler Labine	

	actor_3_name	actor_1_facebook_likes	actor_2_facebook_likes	\
0	Amiée Conn	14000	19000.0	
1	Idris Elba	2800	28000.0	
2	Rooney Mara	33000	96000.0	
3	Forest Whitaker	35000	5300.0	
4	Kyle Chandler	518	71000.0	
..	…	…	…	
95	Chris Mulkey	24000	970.0	
96	Athina Rachel Tsangari	140	63.0	
97	0	131	12.0	
98	Alfred Enoch	10000	2000.0	
99	Chelan Simmons	948	779.0	

	actor_3_facebook_likes	…	Votes3044M	Votes3044F	Votes45A	Votes45AM	\
0	NaN	…	7.9	7.8	7.6	7.6	
1	27000.0	…	7.8	8.1	7.8	7.8	
2	9800.0	…	7.9	8.2	8.0	7.9	
3	NaN	…	7.8	7.8	7.6	7.6	
4	3300.0	…	7.7	7.7	7.6	7.6	
..	…	…	…	…	…	…	
95	535.0	…	8.3	8.2	8.1	8.1	
96	48.0	…	7.8	7.6	7.3	7.4	
97	0.0	…	7.9	8.2	7.9	7.8	
98	1000.0	…	7.3	8.1	7.4	7.3	
99	440.0	…	7.5	7.7	7.5	7.4	

	Votes45AF	Votes1000	VotesUS	VotesnUS	content_rating	Country
0	7.5	7.1	8.3	8.1	PG-13	USA
1	8.1	7.6	8.0	8.0	PG	USA
2	8.4	7.1	8.1	8.0	PG-13	Australia
3	7.7	7.3	8.0	7.9	PG-13	USA
4	7.6	7.1	7.9	7.8	R	USA
..	…	…	…	…	…	…
95	8.2	8.0	8.6	8.4	R	USA

```
96        7.2         7.0       8.0       7.9             R       USA
97        8.2         7.7       8.2       7.9         PG-13       USA
98        8.0         6.7       7.9       7.5         PG-13        UK
99        7.7         7.1       7.7       7.5             R    Canada

[100 rows x 62 columns]
```

[22]:
```python
# Remove numeric null values
numeric_columns = data.select_dtypes(include=[np.number])
data_cleaned = data.dropna(subset=numeric_columns.columns)

# Replace character null values with 'NAN'
data_cleaned = data_cleaned.fillna('NAN')
```

[23]:
```python
data
```

[23]:
```
                                             Title  title_year     budget  \
0                                       La La Land        2016   30000000
1                                         Zootopia        2016  150000000
2                                             Lion        2016   12000000
3                                          Arrival        2016   47000000
4                              Manchester by the Sea        2016    9000000
..                                             …           …          …
95                                        Whiplash        2014    3300000
96                                 Before Midnight        2013    3000000
97    Star Wars: Episode VII - The Force Awakens        2015  245000000
98  Harry Potter and the Deathly Hallows: Part I        2010  150000000
99                           Tucker and Dale vs Evil        2010    5000000

        Gross             actor_1_name       actor_2_name  \
0   151101803              Ryan Gosling         Emma Stone
1   341268248          Ginnifer Goodwin      Jason Bateman
2    51738905                Dev Patel      Nicole Kidman
3   100546139                Amy Adams      Jeremy Renner
4    47695371            Casey Affleck  Michelle Williams
..        …                     …                 …
95   13092000             J.K. Simmons    Melissa Benoist
96    8114507  Seamus Davey-Fitzpatrick       Ariane Labed
97  936662225               Doug Walker         Rob Walker
98  296347721              Rupert Grint         Toby Jones
99     223838           Katrina Bowden        Tyler Labine

         actor_3_name  actor_1_facebook_likes  actor_2_facebook_likes  \
0          Amiée Conn                   14000                 19000.0
1          Idris Elba                    2800                 28000.0
2         Rooney Mara                   33000                 96000.0
3      Forest Whitaker                  35000                  5300.0
```

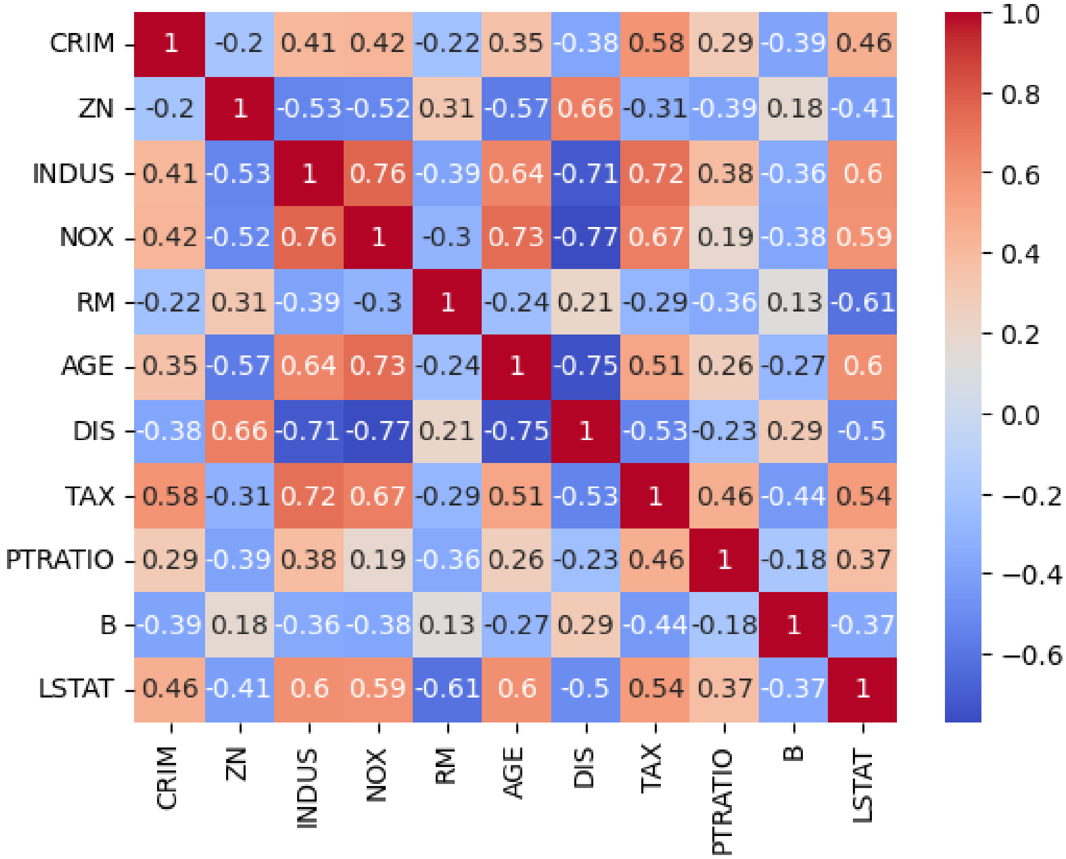
```
4           Kyle Chandler                        518              71000.0
..                      …                          …                    …
95           Chris Mulkey                      24000                970.0
96   Athina Rachel Tsangari                      140                 63.0
97                      0                        131                 12.0
98           Alfred Enoch                      10000               2000.0
99         Chelan Simmons                        948                779.0

    actor_3_facebook_likes  …  Votes3044M Votes3044F Votes45A Votes45AM  \
0                      NaN  …         7.9        7.8      7.6       7.6
1                  27000.0  …         7.8        8.1      7.8       7.8
2                   9800.0  …         7.9        8.2      8.0       7.9
3                      NaN  …         7.8        7.8      7.6       7.6
4                   3300.0  …         7.7        7.7      7.6       7.6
..                     …  …         …        …      …       …
95                   535.0  …         8.3        8.2      8.1       8.1
96                    48.0  …         7.8        7.6      7.3       7.4
97                     0.0  …         7.9        8.2      7.9       7.8
98                  1000.0  …         7.3        8.1      7.4       7.3
99                   440.0  …         7.5        7.7      7.5       7.4

    Votes45AF  Votes1000  VotesUS  VotesnUS  content_rating    Country
0         7.5        7.1      8.3       8.1           PG-13        USA
1         8.1        7.6      8.0       8.0              PG        USA
2         8.4        7.1      8.1       8.0           PG-13  Australia
3         7.7        7.3      8.0       7.9           PG-13        USA
4         7.6        7.1      7.9       7.8               R        USA
..        …        …      …       …             …        …
95        8.2        8.0      8.6       8.4               R        USA
96        7.2        7.0      8.0       7.9               R        USA
97        8.2        7.7      8.2       7.9           PG-13        USA
98        8.0        6.7      7.9       7.5           PG-13         UK
99        7.7        7.1      7.7       7.5               R     Canada

[100 rows x 62 columns]
```

1.2 2.Perform Heatmap using the inbuilt Boston dataset and describe the variable correlation and describe its variate type.

[24]:
```python
import seaborn as sns
import pandas as pd
from sklearn.datasets import fetch_openml

boston = fetch_openml(name='boston', version=1)

# Convert the dataset to a pandas dataframe
df = pd.DataFrame(boston.data, columns=boston.feature_names)
```

```python
# Calculate the correlation matrix
corr_matrix = df.corr()

# Plot the heatmap
sns.heatmap(corr_matrix, cmap='coolwarm', annot=True)
```

[24]: <Axes: >



1.3 3.Import dataset from csv, excel, text and from the google drive link given below LINK (https://drive.google.com/uc?export=download&id=1lqNKpOTdf5va7sQ(I)

1.3.1 a). From excel

[27]:
```python
df = pd.read_excel('EDA_census.xlsx')
df
```

[27]:

```
     Unnamed: 0 Unnamed: 1 Unnamed: 2                   Unnamed: 3  \
0         Table      State     Distt.                    Area Name
1          Name       Code       Code                          NaN
2           NaN        NaN        NaN                          NaN
3           NaN        NaN        NaN                          NaN
4           NaN        NaN        NaN                          NaN
…           …        …        …                          …
3133      C2308         35        000  State - ANDAMAN & NICOBAR ISLANDS
3134      C2308         35        000  State - ANDAMAN & NICOBAR ISLANDS
3135      C2308         35        000  State - ANDAMAN & NICOBAR ISLANDS
3136      C2308         35        000  State - ANDAMAN & NICOBAR ISLANDS
3137      C2308         35        000  State - ANDAMAN & NICOBAR ISLANDS

     Unnamed: 4  \
0        Total/
1        Rural/
2        Urban/
3           NaN
4           NaN
…           …
3134      Urban
3133      Urban
3135      Urban
3136      Urban
3137      Urban

     C-8  EDUCATIONAL LEVEL BY AGE AND SEX FOR POPULATION AGE 7 AND ABOVE - 2011
\
0                                            Age-group
1                                                  NaN
2                                                  NaN
3                                                  NaN
4                                                    1
…                                                  …
3133                                             65-69
3134                                             70-74
3135                                             75-79
3136                                               80+
3137                                    Age not stated

     Unnamed: 6 Unnamed: 7 Unnamed: 8  Unnamed: 9  … Unnamed: 35  \
0         Total        NaN        NaN  Illiterate  …         NaN
1           NaN        NaN        NaN         NaN  …         NaN
2           NaN        NaN        NaN         NaN  …         NaN
3       Persons      Males    Females     Persons  …     Females
4             2          3          4           5  …          31
…           …        …        …        …  …           …
```

```
3133       1757         953         804         586  …          0
3134       1193         691         502         419  …          4
3135        645         343         302         234  …          0
3136        616         287         329         264  …          0
3137        188          94          94          36  …          0

                           Unnamed: 36 Unnamed: 37 Unnamed: 38  \
0                                  NaN         NaN         NaN
1     Technical diploma or certificate         NaN         NaN
2                 not equal to degree          NaN         NaN
3                              Persons       Males     Females
4                                   32          33          34
…                               …         …         …
3133                                21          14           7
3134                                14           9           5
3135                                 6           3           3
3136                                 2           1           1
3137                                 4           2           2

           Unnamed: 39 Unnamed: 40 Unnamed: 41   Unnamed: 42 Unnamed: 43  \
0                  NaN         NaN         NaN           NaN         NaN
1     Graduate & above         NaN         NaN  Unclassified         NaN
2                  NaN         NaN         NaN           NaN         NaN
3              Persons       Males     Females       Persons       Males
4                   35          36          37            38          39
…                …         …         …          …         …
3133                82          72          10             4           1
3134                56          41          15             7           4
3135                22          16           6             7           4
3136                23          14           9             6           1
3137                14           7           7             3           1

     Unnamed: 44
0            NaN
1            NaN
2            NaN
3        Females
4             40
…            …
3133           3
3134           3
3135           3
3136           5
3137           2

[3138 rows x 45 columns]
```

1.3.2 b). From text

[28]:
```
with open('EDA.txt', 'r') as file:
    for line in file:
        print(line)
```

```
In statistics, exploratory data analysis is an approach of analyzing data sets
to summarize their main characteristics, often using statistical graphics and
other data visualization methods.
```

1.3.3 c). From Google Drive

[25]:
```
!pip install gdown

import gdown

url = 'https://drive.google.com/uc?
 ↪export=download&id=1lqNKpOTdf5va7sQOsJKLShIWBvighaAI'
output = 'data.csv'
gdown.download(url, output, quiet=False)
```

```
Collecting gdown
  Downloading gdown-4.7.1-py3-none-any.whl (15 kB)
Requirement already satisfied: requests[socks] in
c:\users\lenovo\anaconda3\lib\site-packages (from gdown) (2.29.0)
Requirement already satisfied: beautifulsoup4 in
c:\users\lenovo\anaconda3\lib\site-packages (from gdown) (4.12.2)
Requirement already satisfied: six in c:\users\lenovo\anaconda3\lib\site-
packages (from gdown) (1.16.0)
Requirement already satisfied: tqdm in c:\users\lenovo\anaconda3\lib\site-
packages (from gdown) (4.65.0)
Requirement already satisfied: filelock in c:\users\lenovo\anaconda3\lib\site-
packages (from gdown) (3.9.0)
Requirement already satisfied: soupsieve>1.2 in
c:\users\lenovo\anaconda3\lib\site-packages (from beautifulsoup4->gdown) (2.4)
Requirement already satisfied: charset-normalizer<4,>=2 in
c:\users\lenovo\anaconda3\lib\site-packages (from requests[socks]->gdown)
(2.0.4)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in
c:\users\lenovo\anaconda3\lib\site-packages (from requests[socks]->gdown)
(1.26.16)
Requirement already satisfied: certifi>=2017.4.17 in
c:\users\lenovo\anaconda3\lib\site-packages (from requests[socks]->gdown)
(2023.5.7)
Requirement already satisfied: idna<4,>=2.5 in
c:\users\lenovo\anaconda3\lib\site-packages (from requests[socks]->gdown) (3.4)
Requirement already satisfied: PySocks!=1.5.7,>=1.5.6 in
```

```
c:\users\lenovo\anaconda3\lib\site-packages (from requests[socks]->gdown)
(1.7.1)
Requirement already satisfied: colorama in c:\users\lenovo\anaconda3\lib\site-
packages (from tqdm->gdown) (0.4.6)
Installing collected packages: gdown
Successfully installed gdown-4.7.1
```

```
Downloading…
From:
https://drive.google.com/uc?export=download&id=1lqNKpOTdf5va7sQOsJKLShIWBvighaAI
To: C:\Users\LENOVO\OneDrive - Lovely Professional University\5th_Sem\INT351
STATISTICS AND EDA\CA\CA1\data.csv
100%|
     | 11.3k/11.3k [00:00<?, ?B/s]
```

[25]: 'data.csv'

1.3.4 d). From CSV

[26]:
```python
df = pd.read_csv("data.csv")
df
```

[26]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	\
0	63	1	3	145	233	1	0	150	0	2.3	0	
1	37	1	2	130	250	0	1	187	0	3.5	0	
2	41	0	1	130	204	0	0	172	0	1.4	2	
3	56	1	1	120	236	0	1	178	0	0.8	2	
4	57	0	0	120	354	0	1	163	1	0.6	2	
..	… …	..	…	… …	…	…	…	… …				
298	57	0	0	140	241	0	1	123	1	0.2	1	
299	45	1	3	110	264	0	1	132	0	1.2	1	
300	68	1	0	144	193	1	1	141	0	3.4	1	
301	57	1	0	130	131	0	1	115	1	1.2	1	
302	57	0	1	130	236	0	0	174	0	0.0	1	

	caa	thall	output
0	0	1	1
1	0	2	1
2	0	2	1
3	0	2	1
4	0	2	1
..	…	…	…
298	0	3	0
299	0	3	0
300	2	3	0
301	1	3	0
302	1	2	0

```
[303 rows x 14 columns]
```

1.4 4. Analyse the data quality issue and give a solution using missing value imputation.

[32]:
```python
# Analyze missing values
missing_values = data.isnull().sum()

# Impute missing values with mean, median, and mode
data_imputed_mean = data.fillna(data.mean())

data_imputed_median = data.fillna(data.median())

data_imputed_mode = data.fillna(data.mode().iloc[0])

# Remove duplicates
data_no_duplicates = data.drop_duplicates()
```

[33]:
```python
data
```

[33]:
```
                                          Title  title_year     budget  \
0                                    La La Land        2016   30000000
1                                      Zootopia        2016  150000000
2                                          Lion        2016   12000000
3                                       Arrival        2016   47000000
4                          Manchester by the Sea        2016    9000000
..                                          …           …          …
95                                     Whiplash        2014    3300000
96                              Before Midnight        2013    3000000
97    Star Wars: Episode VII - The Force Awakens        2015  245000000
98  Harry Potter and the Deathly Hallows: Part I        2010  150000000
99                        Tucker and Dale vs Evil        2010    5000000

        Gross               actor_1_name       actor_2_name  \
0   151101803               Ryan Gosling         Emma Stone
1   341268248            Ginnifer Goodwin      Jason Bateman
2    51738905                  Dev Patel      Nicole Kidman
3   100546139                 Amy Adams      Jeremy Renner
4    47695371              Casey Affleck  Michelle Williams
..        …                        …                  …
95   13092000               J.K. Simmons    Melissa Benoist
96    8114507   Seamus Davey-Fitzpatrick       Ariane Labed
97  936662225                Doug Walker         Rob Walker
98  296347721               Rupert Grint         Toby Jones
99     223838             Katrina Bowden       Tyler Labine

              actor_3_name  actor_1_facebook_likes  actor_2_facebook_likes  \
```

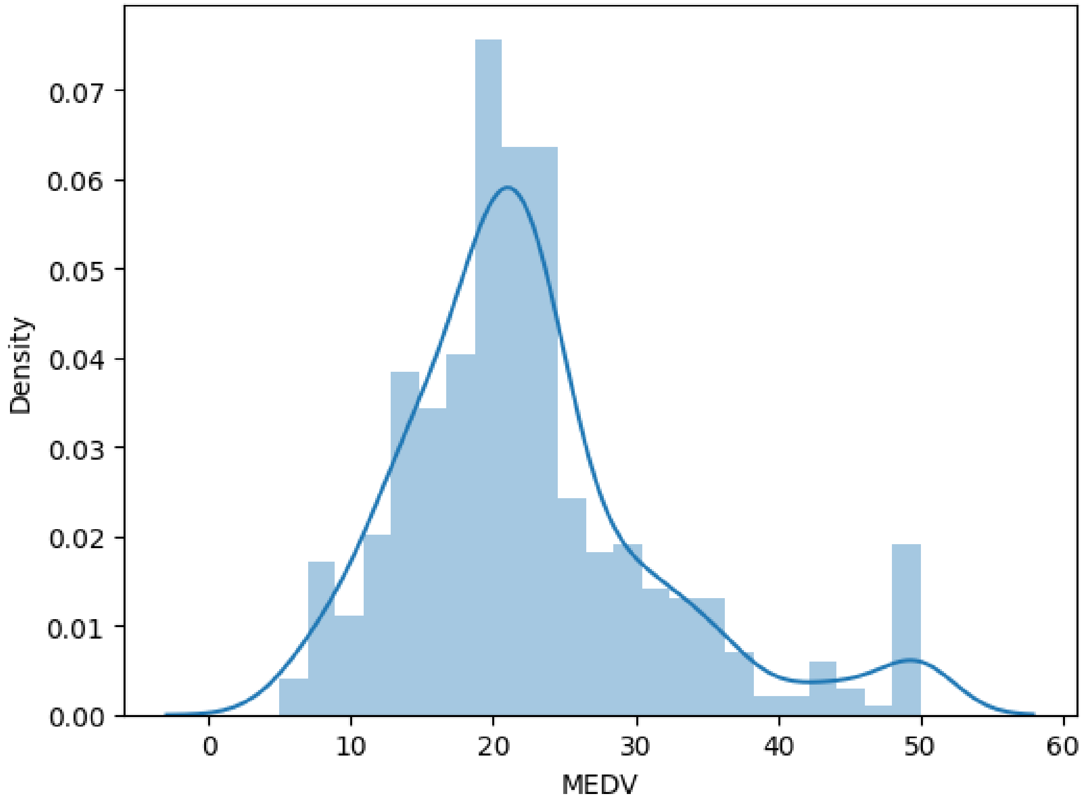
```
0                Amiée Conn                14000              19000.0
1                Idris Elba                 2800              28000.0
2              Rooney Mara                 33000              96000.0
3          Forest Whitaker                 35000               5300.0
4            Kyle Chandler                   518              71000.0
..                       …                     …                    …
95            Chris Mulkey                 24000                970.0
96  Athina Rachel Tsangari                   140                 63.0
97                       0                   131                 12.0
98            Alfred Enoch                 10000               2000.0
99          Chelan Simmons                   948                779.0

    actor_3_facebook_likes  …  Votes3044M Votes3044F Votes45A Votes45AM  \
0                      NaN  …         7.9        7.8      7.6       7.6
1                  27000.0  …         7.8        8.1      7.8       7.8
2                   9800.0  …         7.9        8.2      8.0       7.9
3                      NaN  …         7.8        7.8      7.6       7.6
4                   3300.0  …         7.7        7.7      7.6       7.6
..                     …  …         …        …      …       …
95                   535.0  …         8.3        8.2      8.1       8.1
96                    48.0  …         7.8        7.6      7.3       7.4
97                     0.0  …         7.9        8.2      7.9       7.8
98                  1000.0  …         7.3        8.1      7.4       7.3
99                   440.0  …         7.5        7.7      7.5       7.4

    Votes45AF  Votes1000  VotesUS  VotesnUS  content_rating    Country
0         7.5        7.1      8.3       8.1           PG-13        USA
1         8.1        7.6      8.0       8.0              PG        USA
2         8.4        7.1      8.1       8.0           PG-13  Australia
3         7.7        7.3      8.0       7.9           PG-13        USA
4         7.6        7.1      7.9       7.8               R        USA
..        …        …      …       …             …        …
95        8.2        8.0      8.6       8.4               R        USA
96        7.2        7.0      8.0       7.9               R        USA
97        8.2        7.7      8.2       7.9           PG-13        USA
98        8.0        6.7      7.9       7.5           PG-13         UK
99        7.7        7.1      7.7       7.5               R     Canada

[100 rows x 62 columns]
```

1.5 5. Perform distplot in target variable using the inbuilt Boston dataset and answer whether is randomly distributed or not.

[37]:
```python
import seaborn as sns
from sklearn.datasets import fetch_openml

boston = fetch_openml(name='boston', version=1)
```

```
# Plot a distplot of the target variable
sns.distplot(boston.target)
```

[37]: <Axes: xlabel='MEDV', ylabel='Density'>



The disribution show by graph is not symmetric, i.e the dataset is randomly distributed

1.6 6. Replace missing value in the bank dataset using mean, median, mode and remove the duplicate

[38]:
```
import pandas as pd

# Load the dataset
df = pd.read_csv('bank_data.csv')

# Replace missing values with mean
df.fillna(df.mean(), inplace=True)

# Replace missing values with median
df.fillna(df.median(), inplace=True)
```

```
# Replace missing values with mode
df.fillna(df.mode().iloc[0], inplace=True)

# Identify and remove duplicate rows
df.drop_duplicates(inplace=True)
```

[39]:
```
df
```

[39]:
```
           banking marketing Unnamed: 1                  Unnamed: 2  \
0      customer id and age.          32  Customer salary and balance.
1                customerid         age                       salary
2                         1          58                       100000
3                         2          44                        60000
4                         3          33                       120000
…                         …         …                            …
45208                 45207        51.0                        60000
45209                 45208        71.0                        55000
45210                 45209        72.0                        55000
45211                 45210        57.0                        20000
45212                 45211        37.0                       120000

      Unnamed: 3                                      Unnamed: 4  \
0              0  Customer marital status and job with education…
1        balance                                         marital
2           2143                                         married
3             29                                          single
4              2                                         married
…            …                                                 …
45208        825                                         married
45209       1729                                        divorced
45210       5715                                         married
45211        668                                         married
45212       2971                                         married

                     Unnamed: 5                           Unnamed: 6  \
0          management,tertiary  particular customer before targeted or not
1                       jobedu                             targeted
2          management,tertiary                                  yes
3         technician,secondary                                  yes
4       entrepreneur,secondary                                  yes
…                            …                                    …
45208      technician,tertiary                                  yes
45209          retired,primary                                  yes
45210        retired,secondary                                  yes
45211     blue-collar,secondary                                  yes
45212   entrepreneur,secondary                                  yes
```

```
      Unnamed: 7                         Unnamed: 8 Unnamed: 9   Unnamed: 10  \
0             no  Loan types: loans or housing loans         no  Contact type
1        default                            housing       loan       contact
2             no                                yes         no       unknown
3             no                                yes         no       unknown
4             no                                yes        yes       unknown
…            …                                  …          …           …
45208         no                                 no         no      cellular
45209         no                                 no         no      cellular
45210         no                                 no         no      cellular
45211         no                                 no         no     telephone
45212         no                                 no         no      cellular

      Unnamed: 11       Unnamed: 12          Unnamed: 13 Unnamed: 14  \
0              20  month of contact      duration of call           1
1             day             month              duration    campaign
2               5         may, 2017               261 sec           1
3               5         may, 2017               151 sec           1
4               5         may, 2017                76 sec           1
…              …                 …                    …          …
45208          17         nov, 2017  16.2833333333333 min           3
45209          17         nov, 2017               7.6 min           2
45210          17         nov, 2017  18.7833333333333 min           5
45211          17         nov, 2017  8.46666666666667 min           4
45212          17         nov, 2017  6.01666666666667 min           2

      Unnamed: 15 Unnamed: 16                 Unnamed: 17  \
0              -1           0  outcome of previous contact
1           pdays    previous                     poutcome
2              -1           0                      unknown
3              -1           0                      unknown
4              -1           0                      unknown
…              …          …                         …
45208          -1           0                      unknown
45209          -1           0                      unknown
45210         184           3                      success
45211          -1           0                      unknown
45212         188          11                        other

                               Unnamed: 18
0      response of customer after call happned
1                                 response
2                                       no
3                                       no
4                                       no
…                                      …
45208                                  yes
```

```
45209                                          yes
45210                                          yes
45211                                           no
45212                                           no

[45213 rows x 19 columns]
```

```
[ ]:
```