Q1) Identify the Data type for the Following:

Activity	Data Type
Number of beatings from Wife	Discrete
Results of rolling a dice	Discrete
Weight of a person	Continuous
Weight of Gold	Continuous
Distance between two places	Continuous
Length of a leaf	Continuous
Dog's weight	Continuous
Blue Color	Discrete
Number of kids	Discrete
Number of tickets in Indian railways	Discrete
Number of times married	Discrete
Gender (Male or Female)	Discrete

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

Data	Data Type
Gender	Nominal
High School Class Ranking	Ordinal
Celsius Temperature	Interval
Weight	Ratio
Hair Color	Nominal
Socioeconomic Status	Ordinal
Fahrenheit Temperature	Interval
Height	Ratio
Type of living accommodation	Ordinal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Interval
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Interval
Time on a Clock with Hands	Interval
Number of Children	Ratio
Religious Preference	Nominal
Barometer Pressure	Interval

SAT Scores	Interval
Years of Education	Ratio

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Total number of events (S)= {HHH, HHT, HTT, TTT, TTH, THH, HTH, THT}

n(S)=8

E= {HHT, THH, HTH}

n(E)=3

P(E)=n(E)/n(S)

P(E)=3/8= 0.375

Q4) Two Dice are rolled, find the probability that sum is

a) Equal to 1
b) Less than or equal to 4
c) Sum is divisible by 2 and 3

Ans: Total Numbers of outcome =6*6=36

S=(1, 1)(1, 2)(1, 3)(1, 4)(1, 5)(1, 6) (2, 1)(2, 2)(2, 3)(2, 4)(2, 5)(2, 6) (3, 1) (3, 2)(3,3)(3, 4)(3, 5)(3, 6) (4, 1)(4, 2)(4, 3)(4, 4)(4, 5)(4, 6)(5, 1)(5, 2)(5, 3) (5, 4)(5, 5)(5, 6)(6, 1)(6, 2)(6, 3)(6, 4)(6, 5)(6, 6)
N(S)=36

a) Equal to 1= 0% probability
b) Less than or equal to 4= 6/36 = 1/6
c) Sum is divisible by 2 and 3=

={2 3 4 5 6 7 3 4 5 6 7 8 4 5 6
7 8 9 5 6 7 8 9 10 6 7 8 9
10 11 7 8 9 10 11 12}
=5/36

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

Ans=

Total no. of balls= 2 red +3 green +2 blue= 7

S= two balls are drawn randomly

N(S)= nC_r =$7C_2$ = 7 * 6 /2 * 1 = 21

A= none of the ball drawn is blue

N(A)= $5C_2$= 5 *4 / 2 *1= 10

P(A)= n(A)/n(S) =10/21 = 0.47

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

CHILD	Candies count	Probability
A	1	0.015
B	4	0.20
C	3	0.65
D	5	0.005
E	6	0.01
F	2	0.120

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Ans= Expected number = E(x) =μ_x=1*0.015+4*0.20+3*0.65+5*0.005+6*0.01+2*0.120= **3.09**

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

- For Points,Score,Weigh>
  Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

  ANS=Using Python

In [2]:
```
from scipy import stats
import pandas
```

In [5]:
```
data = pandas.read_csv("Q7.csv")
data.head(5)
```

Out[5]:

	Unnamed: 0	Points	Score	Weigh
0	Mazda RX4	3.90	2.620	16.46
1	Mazda RX4 Wag	3.90	2.875	17.02
2	Datsun 710	3.85	2.320	18.61
3	Hornet 4 Drive	3.08	3.215	19.44
4	Hornet Sportabout	3.15	3.440	17.02

In [6]:
```
data.describe()
```

Out[6]:

	Points	Score	Weigh
count	32.000000	32.000000	32.000000
mean	3.596563	3.217250	17.848750
std	0.534679	0.978457	1.786943
min	2.760000	1.513000	14.500000
25%	3.080000	2.581250	16.892500
50%	3.695000	3.325000	17.710000
75%	3.920000	3.610000	18.900000
max	4.930000	5.424000	22.900000

Points

```
In [4]: data["Points"].mean()
Out[4]: 3.5965625000000006

In [5]: data["Points"].median()
Out[5]: 3.6950000000000003

In [6]: data["Points"].mode()
Out[6]: 0    3.07
        1    3.92
        dtype: float64

In [7]: data["Points"].var()
Out[7]: 0.28588135080645166

In [8]: data["Points"].std()
Out[8]: 0.5346787360709716

In [9]: np.ptp(data["Points"])     #RAGNE
Out[9]: 2.17
```

Score

```
In [10]: data["Score"].mean()
Out[10]: 3.2172499999999995

In [11]: data["Score"].median()
Out[11]: 3.325

In [12]: data["Score"].mode()
Out[12]: 0    3.44
         dtype: float64

In [13]: data["Score"].var()
Out[13]: 0.9573789677419356

In [14]: data["Score"].std()
Out[14]: 0.9784574429896967

In [15]: np.ptp(data["Score"]) #RANGE
Out[15]: 3.9110000000000005
```

Weigh

```
In [17]: data["Weigh"].mean()
Out[17]: 17.848750000000003

In [18]: data["Weigh"].median()
Out[18]: 17.71

In [19]: data["Weigh"].mode()
Out[19]: 0    17.02
         1    18.90
         dtype: float64

In [20]: data["Weigh"].var()
Out[20]: 3.193166129032258

In [21]: data["Weigh"].std()
Out[21]: 1.7869432360968431

In [16]: np.ptp(data["Weigh"]) # RANGE
Out[16]: 8.399999999999999
```

Q8) Calculate Expected Value for the problem below

a) The weights (X) of patients at a clinic (in pounds), are
108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

ANS:

$$EV=\Sigma x/n = \frac{108+110+123+134+135+145+167+187+199}{9} = \underline{145.33}$$

Q9) Calculate Skewness, Kurtosis & draw inferences on the following data

Cars speed and distance

ANS= Using python

In [31]:
```python
import pandas as pd
import numpy as np
data = pd.read_csv("Q9_a.csv")

```

Speed

In [32]:
```python
pandas.Series.skew(data["speed"])
```
Out[32]: -0.11750986144663393

In [33]:
```python
pandas.Series.kurt(data["speed"])
```
Out[33]: -0.5089944204057617

Distance

In [35]:
```python
pandas.Series.skew(data["dist"])
```
Out[35]: 0.8068949601674215

In [36]:
```python
pandas.Series.kurt(data["dist"])
```
Out[36]: 0.4050525816795765

In [37]:
```python
import pandas as pd
import numpy as np
data = pd.read_csv("Q9_b.csv")
```

SP

In [38]:
```python
pandas.Series.skew(data["SP"])
```
Out[38]: 1.6114501961773586

In [39]:
```python
pandas.Series.skew(data["SP"])
```
Out[39]: 2.9773289437871835

WT

In [40]:
```python
pandas.Series.skew(data["WT"])
```
Out[40]: -0.6147533255357768

In [41]:
```python
pandas.Series.skew(data["WT"])
```
Out[41]: -0.6147533255357768

Inferences:

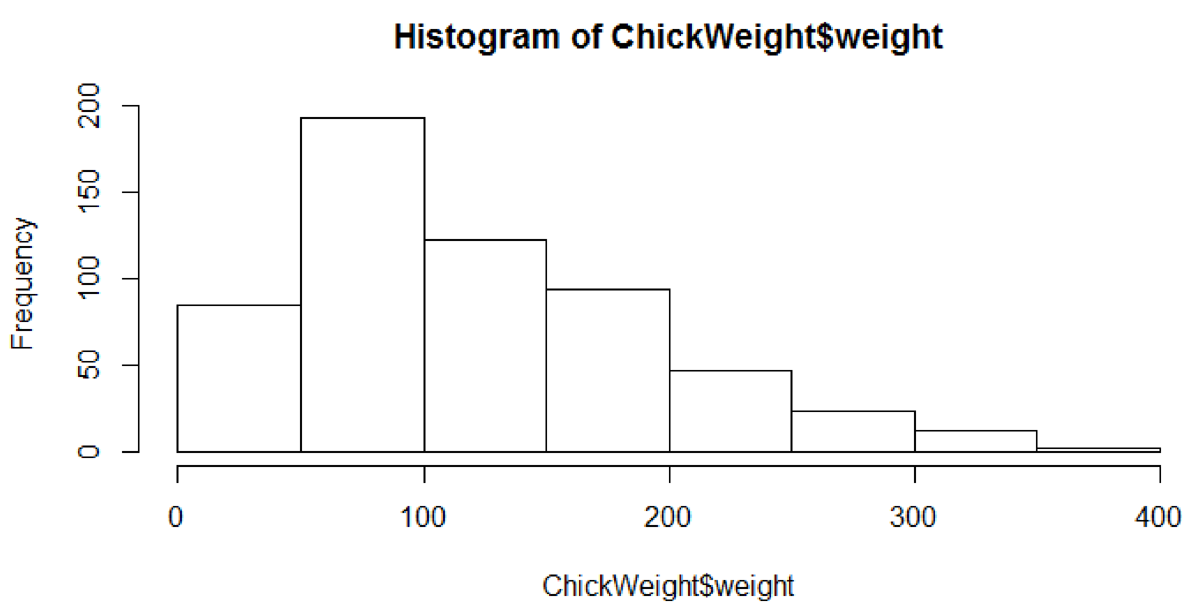
Speed:

1. Skewness = -0.117
   - data is slightly Negatively Skewed or Left Skewed data (Mass of data is on right side of median),
   - means data spared is More on left side of the Median
2. Kurtosis = -0.508
   - Data has platykurtic distribution& has thin tails compared to Normal dist.,
   - The distribution is flat as compared to Normal distribution.

Distance:

1. Skewness = 0.806
   - data is skewed Positively or Right skewed data (Mass of data is on left side of median),
   - Means data spared is more on right side of the Median
2. Kurtosis = 0.405
   - Data has Leptokurtic distribution & has thick tails as compared to normal dist.,The distribution is peak as compared to Normal Distribution.

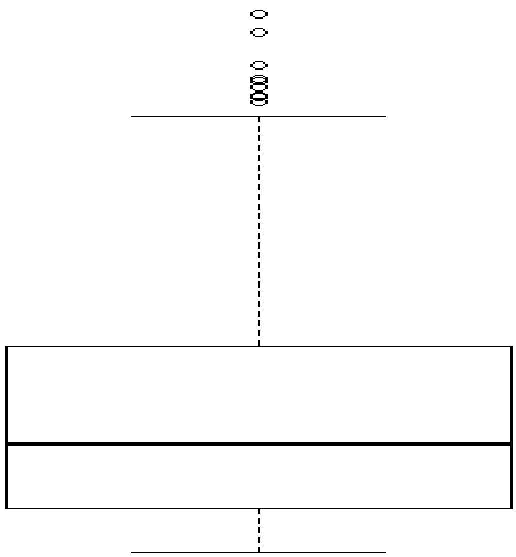
Q10) Draw inferences about the following boxplot & histogram



Inferance :

- Positivelly Skewed data(Right Skewed data)
- Spared of the data on right side of the distribution is More & Mass of data is on left side of Median

Frequency of the data between 50 – 100 is more



Interance:

- Positivelly Skewed data or Right skewed data (Whisker is More on right side on median)
- Spared of the data on right side of the distribution is More & Mass of data is on left side of Median
- Positive Outliers are there on Right side of the distribution

Q11) Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

ANS:

Using Python

```
In [42]:  1 from scipy import stats

In [43]:  1 stats.t.interval(alpha=0.06,df=1999,loc=200,scale=30)
Out[43]: (197.7416201156807, 202.25837988433193)

In [44]:  1 stats.t.interval(alpha=0.04,df=1999,loc=200,scale=30)
Out[44]: (198.49520384079835, 201.50479615920165)

In [45]:  1 stats.t.interval(alpha=0.02,df=1999,loc=200,scale=30)
Out[45]: (199.24783863179837, 200.75216136820163)
```

94%

```
In [5]: stats.t.ppf(q=0.03,df=1999,loc=0,scale=1)
        #ppf(q, df, loc=0, scale=1)
Out[5]: -1.8818614764780115

In [6]: 200-1.8818614764780115*(30/np.sqrt(2000))
Out[6]: 198.7376089443071

In [7]: 200+1.8818614764780115*(30/np.sqrt(2000))
Out[7]: 201.2623910556929
```

96%

```
stats.t.ppf(q=0.02,df=1999,loc=0,scale=1)
#ppf(q, df, loc=0, scale=1)
```

```
-2.055089962825778
```

```
200-2.055089962825778*(30/np.sqrt(2000))
```

```
198.6214037429732
```

```
200+2.055089962825778*(30/np.sqrt(2000))
```

```
201.3785962570268
```

98%

```
stats.t.ppf(q=0.01,df=1999,loc=0,scale=1)
#ppf(q, df, loc=0, scale=1)
```

```
-2.3282147761069725
```

```
200-2.3282147761069725*(30/np.sqrt(2000))
```

```
198.4381860483216
```

```
200+2.3282147761069725*(30/np.sqrt(2000))
```

```
201.5618139516784
```

Q12) Below are the scores obtained by a student in tests

34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

1) Find mean, median, variance, standard deviation.
2) What can we say about the student marks?

1)Ans =

Mean=34+36+36+38+38+39+39+40+40+41+41+41+41+42+42+45+49+56/18

=738/18= 41

=AVERAGE(K3:K20)=41

Median= =MEDIAN(K3:K20)

=40.5

Varience= =VAR.S(K3:K20)

=25.529

Standard Deviation= =stdev.s(K3:K20)

=5.052

2) Mean > Median, This implies that the distribution is slightly skewed towards right. No outliers are present.

Mean = 41 • Most of students' marks are nearer to 41
Median = 40.5 ≅ Mean • There is no too high (like 98,76) & too low marks (like 0,2) (Outliers) present
Standard deviation = 5.05 As mean is approximately equal to median follows Normal distribution, • 1σ = (41-5 =36, 41+5 = 47) • 68% of students are scored between 36 to 47 • 2 σ = (41-10=31,41+10=51) • 95% of students are scored between 41 to 51 • 3SD= (41-15=26,41+16=57) • All most all (99.7%) students are scored between 26 to 57

ANS=

USING PYTHON

```
In [1]: import pandas
        import numpy

In [2]: Marks = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

In [3]: Marks = pandas.Series(Marks)

In [4]: Marks.mean() # MEAN

Out[4]: 41.0

In [5]: Marks.median() # MEDIAN

Out[5]: 40.5

In [6]: Marks.mode() # MODE

Out[6]: 0    41
        dtype: int64

In [7]: Marks.var() # VARIANCE

Out[7]: 25.529411764705884

In [8]: Marks.std() # STANDARD DEVIATION

Out[8]: 5.05266382858645

In [9]: numpy.ptp(Marks) # RANNGE

Out[9]: 22
```

Q13) What is the nature of skewness when mean, median of data are equal?

Ans= **Mean = Median,** we can say data is Normally Distributed.

Q14) What is the nature of skewness when mean > median ?

Ans) **Mean > Median,** we can say Positively Skewed data (Right Skewed data).

Q15) What is the nature of skewness when median > mean?

Ans) **Mean < Median**, we can say Negatively Skewed data (left Skewed data).

Q16) What does positive kurtosis value indicates for a data ?

Ans) Positive Kurtosis (Excess Kurtosis) indicates that,

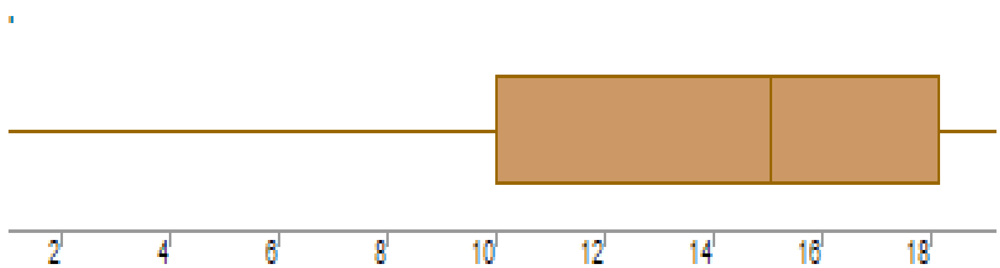
- Distribution is Leptokurtic (peak of bell curve is more as compared to Normal distribution)
- Spared There are more values around mean.

Q17) What does negative kurtosis value indicates for a data?

Ans) Negative Kurtosis indicates that,

- Distribution is Platykurtic (peak of bell curve is less as compared to Normal distribution)
- Spread of the data is More (There are more far values from mean).

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

- Most of the data lies between 10 t0 18.
- Q1 = Quartile 1 = 10
- Q2 = Quartile 2 = 15 = MEDIAN = 50th Percentile

•	Q3 = Quartile 3 = 18

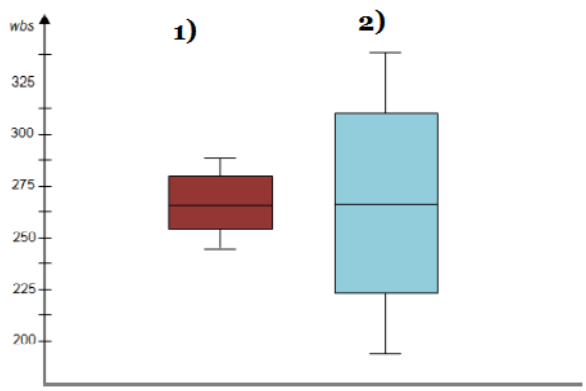
What is nature of skewness of the data?

Negatively skewed data: There are negative outliers present in the data

What will be the IQR of the data (approximately)?

IQR = Q3 – Q1 = 18 -10 = 8

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

Ans)

Boxplot 1	**Boxplot 2**
Data ranges between 240 to 280	Data ranges between 190 to 340
Mean = Median = Mode = Qurtile2(Q2) =260	Mean = Median = Mode= Qurtile2(Q2) =260
Normally Distributed	Normally Distributed
Quartile1 = 255	Quartile1 = 220
Quartile3 = 280	Quartile3 = 310
IQR (INTER QURTAIL RANGE) is less = 280-255 =25	IQR (INTER QURTAIL RANGE) is more = 310-220 =90

Q 20) Calculate probability from the given dataset for the below cases

Data _set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

a. P(MPG>38)
b. P(MPG<40)
c. P (20<MPG<50)

Ans= Using Python

In [1]:
```
from scipy import stats
import pandas
```

In [2]:
```
data = pandas.read_csv("Cars.csv")
data.head(5)
```

Out[2]:

	HP	MPG	VOL	SP	WT
0	49	53.700681	89	104.185353	28.762059
1	55	50.013401	92	105.461264	30.466833
2	55	50.013401	92	105.461264	30.193597
3	70	45.696322	92	113.461264	30.632114
4	53	50.504232	92	104.461264	29.889149

In [3]:
```
MPG=data["MPG"]
MPG.head()
```

Out[3]:
```
0    53.700681
1    50.013401
2    50.013401
3    45.696322
4    50.504232
Name: MPG, dtype: float64
```

a. P(MPG>38)

In [4]:
```
1- stats.norm.cdf(x=38,loc=MPG.mean(),scale=MPG.std())
```

Out[4]:
```
0.3475939251582705
```

b. P(MPG<40)

```
In [5]: stats.norm.cdf(x=40,loc=MPG.mean(),scale=MPG.std())
Out[5]: 0.7293498762151616
```

c. P (20<MPG<50)

```
In [6]: X1_20 = stats.norm.cdf(x=20,loc=MPG.mean(),scale=MPG.std())
        X1_20
Out[6]: 0.05712377632115936

In [7]: X2_58 = stats.norm.cdf(x=50,loc=MPG.mean(),scale=MPG.std())
        X2_58
Out[7]: 0.955992693289364

In [8]: P = X2_58 - X1_20
        P
Out[8]: 0.8988689169682046
```

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

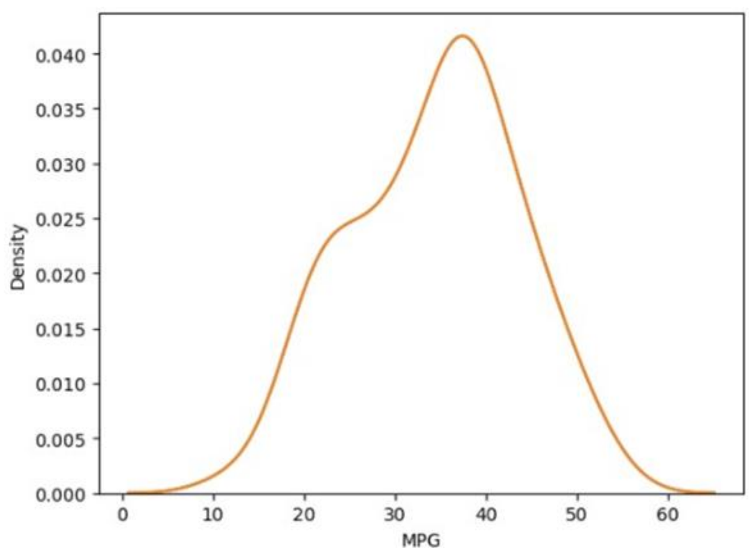
```
In [1]: import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns

In [31]: data = pd.read_csv("Cars.csv")
         data
Out[31]:
```

	HP	MPG	VOL	SP	WT
0	49	53.700681	89	104.185353	28.762059
1	55	50.013401	92	105.461264	30.466833
2	55	50.013401	92	105.461264	30.193597
3	70	45.696322	92	113.461264	30.632114
4	53	50.504232	92	104.461264	29.889149
...	...	...	...	...	...
76	322	36.900000	50	169.598513	16.132947
77	238	19.197888	115	150.576579	37.923113
78	263	34.000000	50	151.598513	15.769625
79	295	19.833733	119	167.944460	39.423099
80	236	12.101263	107	139.840817	34.948615

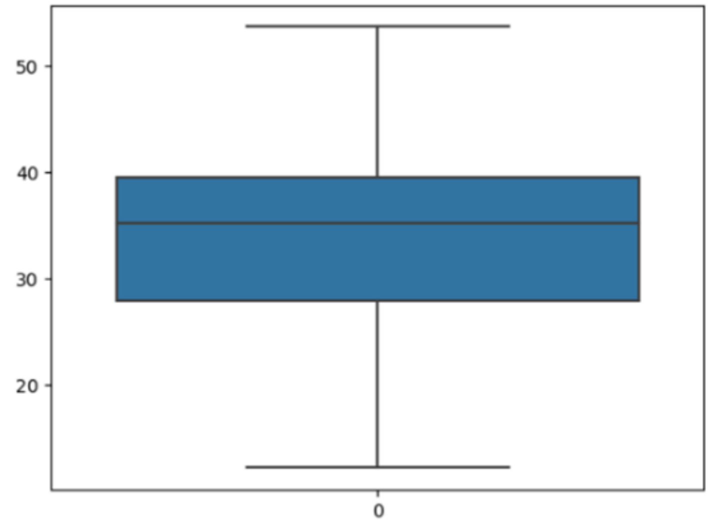
81 rows × 5 columns

```
In [3]: data["MPG"].mean()
Out[3]: 34.422075728024666

In [4]: data["MPG"].median()
Out[4]: 35.15272697

In [5]: data["MPG"].mode()
Out[5]: 0    29.629936
        dtype: float64

In [6]: data["MPG"].skew()
Out[6]: -0.17794674747025727

In [7]: data["MPG"].kurt()
Out[7]: -0.6116786559430913
```

```
sns.kdeplot(data["MPG"])
plt.show()
```



```
In [9]: sns.boxplot(data=data["MPG"])
        plt.show()
```



1)MEAN ≠ MEDIAN,
2)Skewness = 0.177
3)Kurtosis = 0.6116
4)IN Box plot Q2 is not at center, whisker is more negative side , Midian(Q2) is nearer to Q3 and in bell curve skewed towards negative numbers

We can Say That the “MPG” data is Sightlly Right skewed or Negatively Skewed data.

b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
Dataset: wc-at.csv

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
data = pd.read_csv("wc-at.csv")
data
```

Out[2]:

	Waist	AT
0	74.75	25.72
1	72.60	25.89
2	81.80	42.60
3	83.95	42.80
4	74.65	29.84
...	...	...
104	100.10	124.00
105	93.30	62.20
106	101.80	133.00
107	107.90	208.00
108	108.50	208.00

109 rows × 2 columns

Waist

In [3]:
```
data["Waist"].mean()
```

Out[3]:
```
91.90183486238533
```

In [4]:
```
data["Waist"].mean()
```

Out[4]:
```
91.90183486238533
```

In [5]:
```
data["Waist"].mode()
```

Out[5]:
```
0     94.5
1    106.0
2    108.5
dtype: float64
```

In [6]:
```
data["Waist"].skew()
```

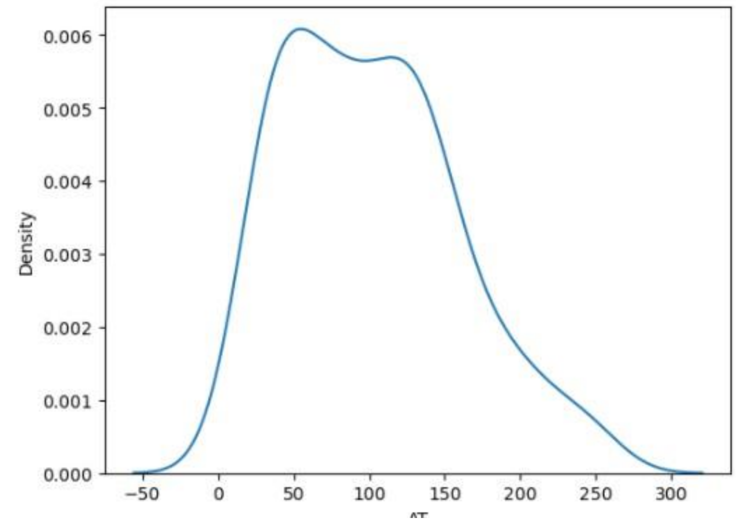
Out[6]:
```
0.1340560824786468
```

In [7]:
```
data["Waist"].kurt()
```

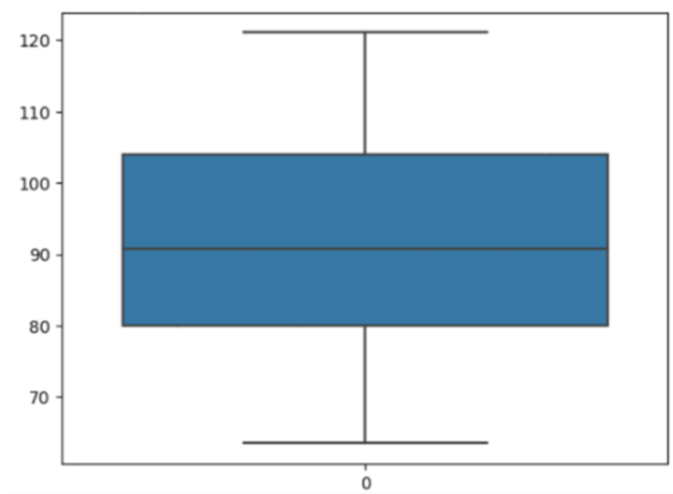
Out[7]:
```
-1.1026666011768886
```

In [8]:
```
sns.kdeplot(data["Waist"])
plt.show()
```

In [14]:
```
sns.kdeplot(data["AT"])
plt.show()
```




```
sns.boxplot(data=data["Waist"])
plt.show()
```



1)MEAN = MEDIAN =91.9018,

2)Skewness = 0.134 ≅ 0

3)Kurtosis = -1.01

4)IN Box plot Q2 is approximately at center

We can Say That the "Waist" data is Normally Distributed

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
data = pd.read_csv("wc-at.csv")
data
```

Out[2]:

	Waist	AT
0	74.75	25.72
1	72.60	25.89
2	81.80	42.60
3	83.95	42.80
4	74.65	29.84
...	...	...
104	100.10	124.00
105	93.30	62.20
106	101.80	133.00
107	107.90	208.00
108	108.50	208.00

109 rows × 2 columns

AT

In [9]:
```
data["AT"].mean()
```

Out[9]: 101.89403669724771

In [10]:
```
data["AT"].median()
```

Out[10]: 96.54

In [11]:
```
data["AT"].mode()
```

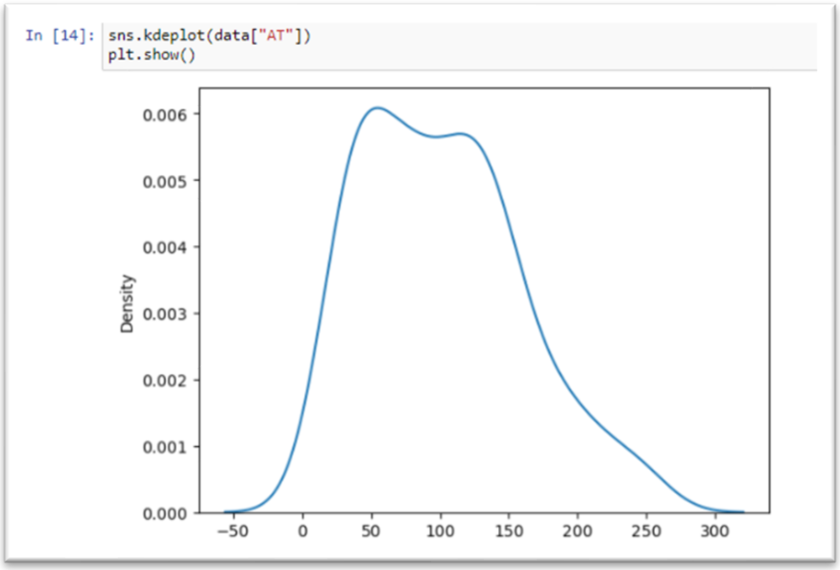
Out[11]:
```
0    121.0
1    123.0
dtype: float64
```

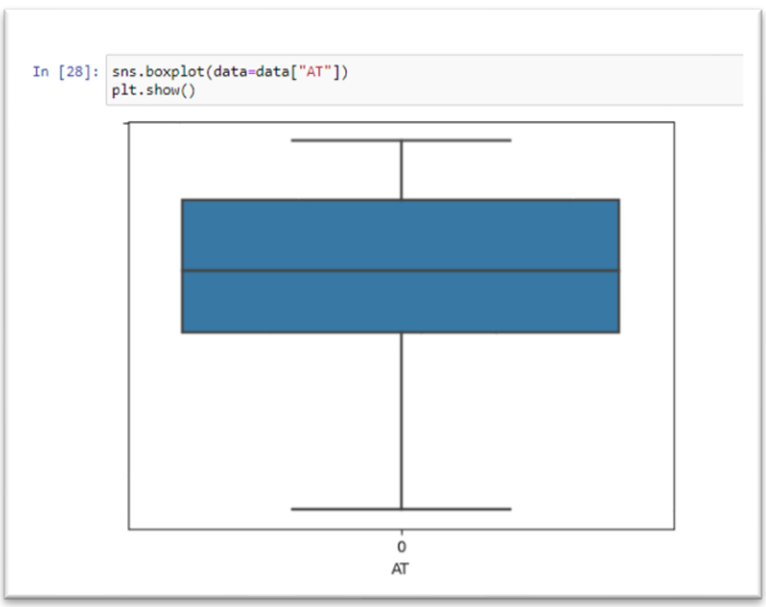
In [12]:
```
data["AT"].skew()
```

Out[12]: 0.584869324127853

In [13]:
```
data["AT"].kurt()
```

Out[13]: -0.28557567504584425





1) MEAN ≠ MEDIAN
2) Skewness, is not nearer zero 3)Kurtosis = -2.855 is not nearer to zero
4)IN Box plot Q2 is not at center and whisker is More in Positive side

We can Say That the “AT” data is Moderately Positively Skewed Data.

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval

ANS=

Confidence Interval	**Alpha(α) =(1-CL)/2**	**Z score (Z table)**
90%	0.10/2=0.05	±1.64
94%	0.06/2=0.03	±1.88
60%	0.40/2=0.20	±0.84

90% →> qnorm(0.95)

[1] 1.644854

94% →> qnorm(0.97)

[1] 1.880794

60% →> qnorm(0.8)

[1] 0.8416212

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

ANS=

Confidence Interval	Df	T score (t table)
95%		2.060
96%	25	2.060
99%		2.787

95% → > qt(0.975,24)

[1] 2.063899

96% → > qt(0.98,24)

[1] 2.171545

99% → qt(0.995,24)

[1] 2.79694

Q 24) A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode → pt(tscore,df)

df → degrees of freedom

ANS=

μ=270

$\bar{x}$=260

SD=90

n=18

df=n-1=18-1= 17

tscore= $\frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{260-270}{90/\sqrt{18}}$ = -10/21.23= -0.47

```
> pt(-0.47,17)
[1] 0.3221639
```

Required probability = 0.32=32%

Ans= Using Python

```
In [1]: import scipy.stats as st

In [2]: st.t.sf(abs(-.4714), 17)

Out[2]: 0.32167411684460556
```