Q1) Identify the Data type for the Following:

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

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

Nominal, Ordinal, Interval, Ratio.

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

Religious Preference	Ordinal
Barometer Pressure	Interval
SAT Scores	Ratio
Years of Education	Ratio

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

Ans;

When three coins are tossed together,

The total number of favorable outcomes = 8

{HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}

Probability = number of favorable outcomes / total number of outcomes

Numbers of outcomes that gives two heads and one tail = 3

{HHT, HTH, THH}

Probability = 3/8 (or) 0.375 =3.75%.

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;

When two dices are rolled n(s) = 6*6 =36

Probability = number of favorable outcomes / total number of outcomes

a) The sum of equal to 1 is 0

There cannot be any probability of 1 outcome

I.e., =0

b) the sum is equal to 4

B= {(1,3), (2,2), (3,1)}

n(B) =n(B)/n(s) =3/36 = 0.0833 =8.33%

C)sum is divisible by both 2 and 3

Favorable outcomes C= {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}

n(C) = n(C)/n(s) = 6/36 =0.166 = 1.66%

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;

Probability = number of favorable outcomes / total number of outcomes

Probability **=** (2R,3G,2B) = (2+3+2) = 7

Total number of outcomes = 7c2

7c2 = (7x6)/(2x1) = 21

Number of favorable outcomes = 5c2

5c2 = (5x4)/(2x1) = 10

Probability = 7c2/5c2

i.e., Probability =10/21= 0.476 = 47.6%

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 random values = $\Sigma X * P(X)$

= 1*0.015 + 4*0.20 + 3*0.65 + 5*0.005 + 6*0.01 + 2*0.120

Expected number of candies for randomly selected child = 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.

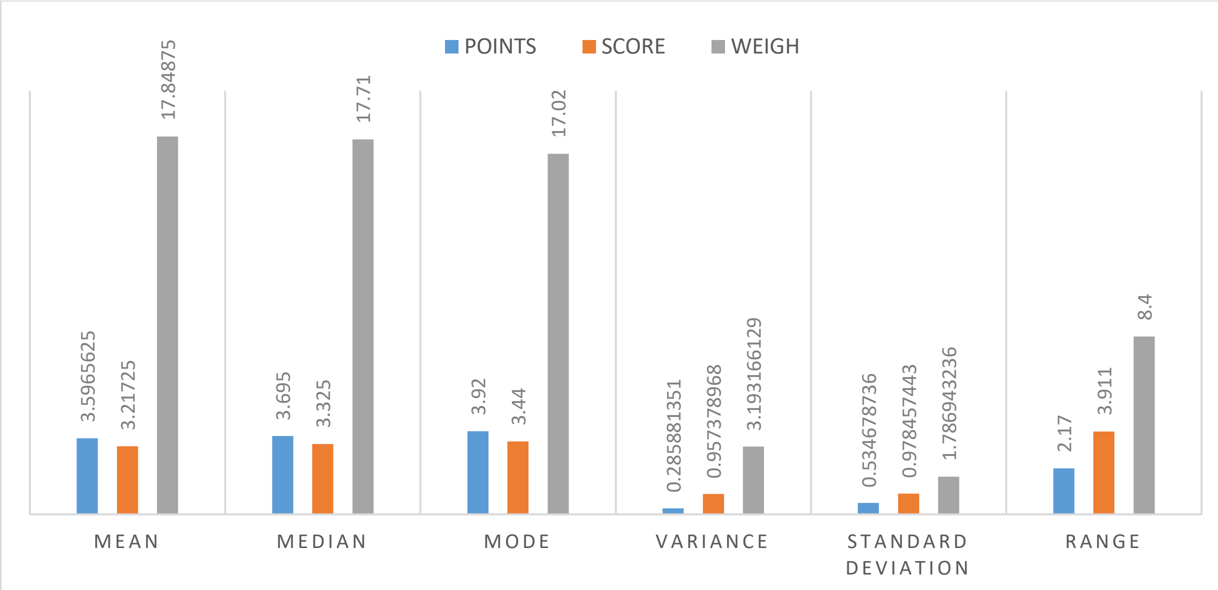
Use Q7.csv file

ANS;

	Points	Score	Weigh
Mean	3.596563	3.21725	17.84875
Median	3.695	3.325	17.71
Mode	3.92	3.44	17.02
Variance	0.285881	0.957379	3.193166
Standard deviation	0.534679	0.978457	1.786943
Range	2.17	3.911	8.4

Inferences;

- Mean value is close for both 'Points' & 'Score'
- Mean = Median =Mode.



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;

Expected value = P(X) * E(X)

Total 9 patients, the probability of each patient P(X)= 1/9

E(X) = 108, 110, 123, 134, 135, 145, 167, 187, 199

Expected value = (1/9) (108+110+123+134+135+145+167+187+199)

= (1/9) (1308)

= 145.33

Expected value of the weight of the patient = 145.33(pounds)

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

Car's speed and distance

Use Q9_a.csv

ANS;

```
In [1]: import pandas as pd
        import numpy as np
```
executed in 941ms, finished 14:09:56 2021-08-30

```
In [2]: df = pd.read_csv("Q9a.csv")
```
executed in 31ms, finished 14:09:57 2021-08-30

```
In [3]: df1=df.iloc[:,1:]
        df1.head()
```
executed in 41ms, finished 14:09:58 2021-08-30

Out[3]:

	speed	dist
0	4	2
1	4	10
2	7	4
3	7	22
4	8	16

```
In [4]: df1.skew()
```
executed in 11ms, finished 14:10:00 2021-08-30

```
Out[4]: speed   -0.117510
        dist     0.806895
        dtype: float64
```

```
In [5]: df1.kurt()
```
executed in 38ms, finished 14:10:01 2021-08-30

```
Out[5]: speed   -0.508994
        dist     0.405053
        dtype: float64
```

	speed	dist.
Skewness	-0.11751	0.806895
Kurtosis	-0.50899	0.405053

Skewness;

- The skewness value for speed is (Negative skewness), so it is left skewed.
- And for distance, is right skewed (positive skewness).

Kurtosis;

- Speed is negative kurtosis, (flatter than normal distribution)
- distance is positive kurtosis (peaked than normal distribution)

SP and Weight (WT)

Use Q9_b.csv

ANS;

In [6]:
```
df2 = pd.read_csv("Q9_b.csv")
```
executed in 28ms, finished 14:10:03 2021-08-30

In [7]:
```
df3=df2.iloc[:,1:]
df3.head()
```
executed in 15ms, finished 14:10:05 2021-08-30

Out[7]:

	SP	WT
0	104.185353	28.762059
1	105.461264	30.466833
2	105.461264	30.193597
3	113.461264	30.632114
4	104.461264	29.889149

In [8]:
```
df3.skew()
```
executed in 38ms, finished 14:10:06 2021-08-30

Out[8]:
```
SP     1.611450
WT    -0.614753
dtype: float64
```

In [9]:
```
df3.kurt()
```
executed in 17ms, finished 14:10:06 2021-08-30

Out[9]:
```
SP     2.977329
WT     0.950291
dtype: float64
```

	SP	WT
Skewness	1.61145	-0.61475
Kurtosis	2.977329	0.950291

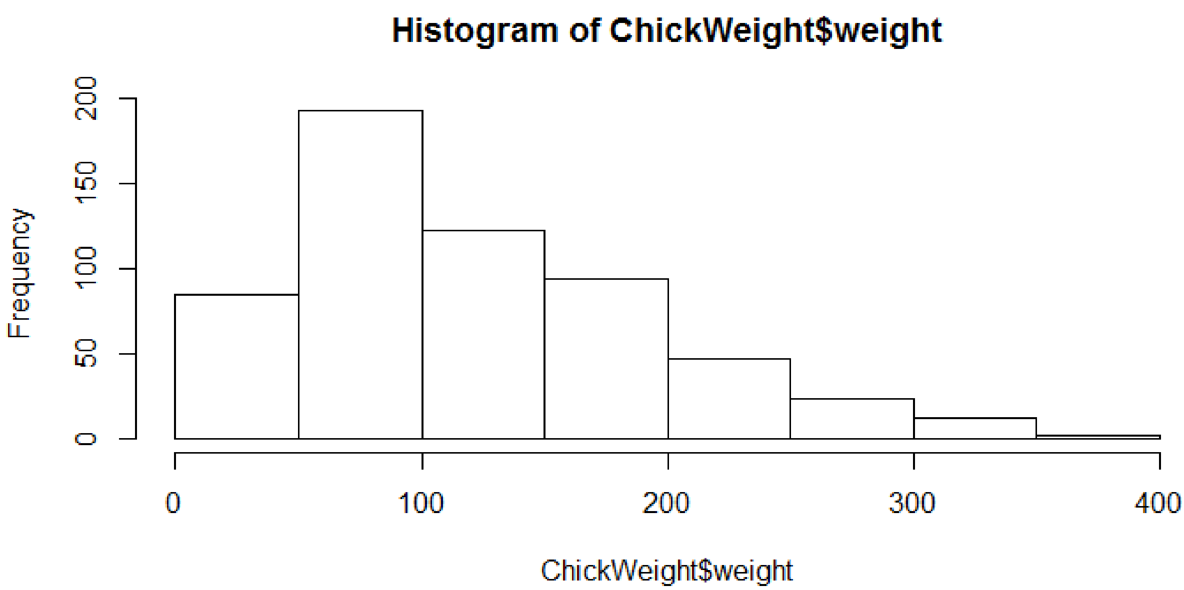
Skewness;

- Sp is right skewed (positive skewness)
- Wt is left skewed (negative skewness)

Kurtosis;

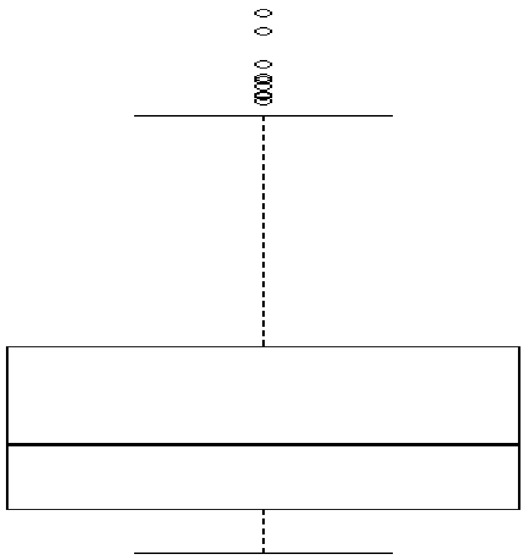
- Peaked than normal distribution

Q10) Draw inferences about the following boxplot & histogram



Ans;

- The most of the data points are concerted in the range 50-100 with in frequency 200.
- And least range of weight is 400 somewhere around 0-10.
- Skewness – we can notice a long tail towards right so it is heavily right skewed.



Ans;

- Median is less than the mean right skewed.
- we have outlier on the upper side of the box plot.
- There are less data points between q1 and bottom points.

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;

In [1]:
```
import pandas as pd
import numpy as np
from scipy import stats
from scipy.stats import norm
```
executed in 885ms, finished 11:31:01 2021-08-28

In [2]:
```
#94%
stats.norm.interval(0.94,200,30/(2000**0.5))
```
executed in 39ms, finished 11:31:02 2021-08-28

Out[2]:
```
(198.738325292158, 201.261674707842)
```

In [3]:
```
#98%
stats.norm.interval(0.98,200,30/(2000**0.5))
```
executed in 30ms, finished 11:31:03 2021-08-28

Out[3]:
```
(198.43943840429978, 201.56056159570022)
```

In [4]:
```
#96%
stats.norm.interval(0.96,200,30/(2000**0.5))
```
executed in 8ms, finished 11:31:04 2021-08-28

Out[4]:
```
(198.62230334813333, 201.37769665186667)
```

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.

Ans;

In [1]:
```
import pandas as pd
import numpy as np
```
executed in 496ms, finished 11:45:51 2021-08-28

In [2]:
```
scores=pd.Series([34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56])
```
executed in 37ms, finished 11:45:52 2021-08-28

In [3]:
```
#mean
scores.mean()
```
executed in 43ms, finished 11:45:53 2021-08-28

Out[3]:
```
41.0
```

In [4]:
```
#median
scores.median()
```
executed in 15ms, finished 11:45:54 2021-08-28

Out[4]:
```
40.5
```

In [5]:
```
#variance
scores.var()
```
executed in 16ms, finished 11:45:54 2021-08-28

Out[5]:
```
25.529411764705884
```

In [6]:
```
#standard deviation
scores.std()
```
executed in 40ms, finished 11:45:56 2021-08-28

Out[6]:
```
5.05266382858645
```

2)What can we say about the student marks?

Ans;

- There are 2 outliers in the student's marks 49 & 56.
- The mean is approximately equal to the median.

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

Ans;

- Mean=median=mode
- perfect skewness (Normally distributed)

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

ANS:

- mean > median
- positively skewed data (Right skewed)

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

ANS:

- mean < median
- negatively skewed data (Left skewness)

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

ANS:

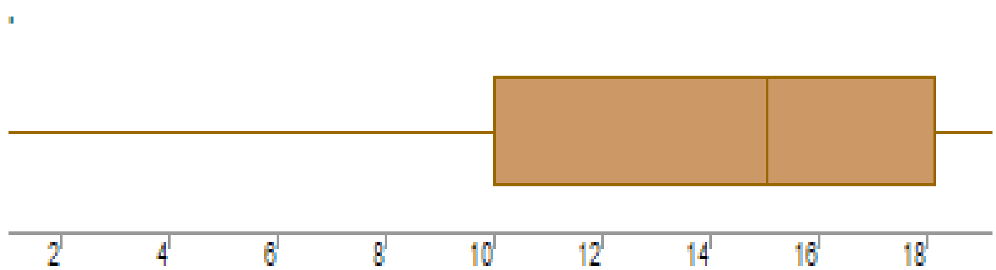
- The data is normally distributed and kurtosis value is 0.
- Bell curve structure.

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

ANS:

- The distribution of the data has lighter tails and a flatter peak than the normal distribution.

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



1. What can we say about the distribution of the data?

ANS;

- Most of the data lies between 10-18
- Quartile
    - Q1 = 10
    - Q2 = 15(MEDIAN)

        - (Most of the values lies below the median.)
    - Q3 = 18
- median is greater than mean.

2. What is nature of skewness of the data?

ANS; Negative skewed data (outliers is present)

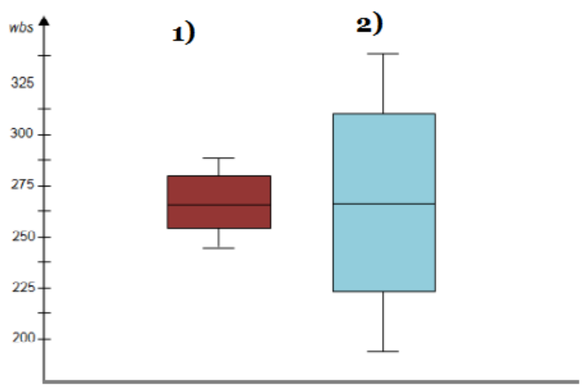
3. What will be the IQR of the data (approximately)?

ANS;

INTER QUARTILE RANGE (IQR) = Q3-Q1 = 18-10

Approximately (IQR) = -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**
Range b/w 240 - 280	Range b/w 190 - 340
Mean = Median = Mode = 260	Mean = Median = Mode = 260

Quartile(Q1) = 255	Quartile(Q1) = 220
Quartile(Q2) = 260	Quartile(Q2) = 260
Quartile(Q3) = 280	Quartile(Q3) = 310
Inter Quartile Range (IQR) = 25	Inter Quartile Range (IQR) = 90

- By observing both the plots whisker's level is high in boxplot 2.
- Mean=median=mode
- perfect skewness (Normally distributed)

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;

In [1]:
```python
import pandas as pd
import numpy as np
from scipy import stats
```
executed in 940ms, finished 10:11:16 2021-08-30

In [2]:
```python
cars = pd.read_csv("cars.csv")
cars.head()
```
executed in 35ms, finished 10:11:17 2021-08-30

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]:
```python
# P(MPG>38)
1-stats.norm.cdf(38,cars.MPG.mean(),cars.MPG.std())
```
executed in 15ms, finished 10:11:18 2021-08-30

Out[3]: 0.3475939251582705

In [4]:
```python
# P(MPG<40)
stats.norm.cdf(40,cars.MPG.mean(),cars.MPG.std())
```
executed in 16ms, finished 10:11:18 2021-08-30

Out[4]: 0.7293498762151616

```
# P(20<MPG<50)
```

In [5]:
```python
X1=stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())
X1
```
executed in 35ms, finished 10:11:19 2021-08-30

Out[5]: 0.05712377632115936

In [6]:
```python
X2=stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std())
X2
```
executed in 36ms, finished 10:11:22 2021-08-30

Out[6]: 0.955992693289364

In [7]:
```python
# P(20<MPG<50)
X=X2-X1
X
```
executed in 15ms, finished 10:11:23 2021-08-30

Out[7]: 0.8988689169682046

Q 21) Check whether the data follows normal distribution

a) Check whether the MPG of Cars follows Normal Distribution
   Dataset: Cars.csv

ANS;

In [1]:
```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```
executed in 1.36s, finished 10:22:43 2021-08-30

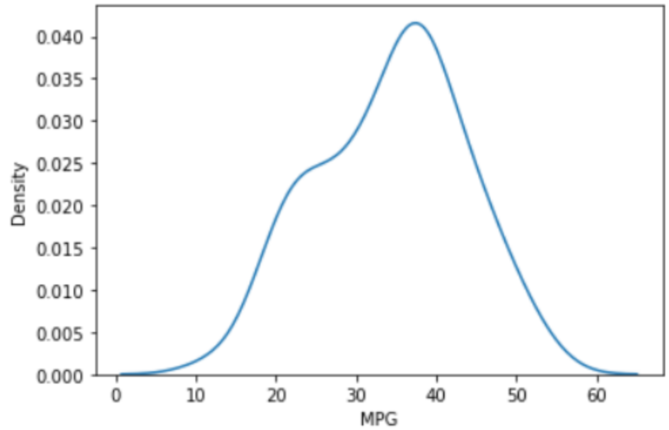
In [2]:
```python
car =pd.read_csv("Cars.csv")
car
```
executed in 48ms, finished 10:22:43 2021-08-30

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
...	...	...	...	...	...
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

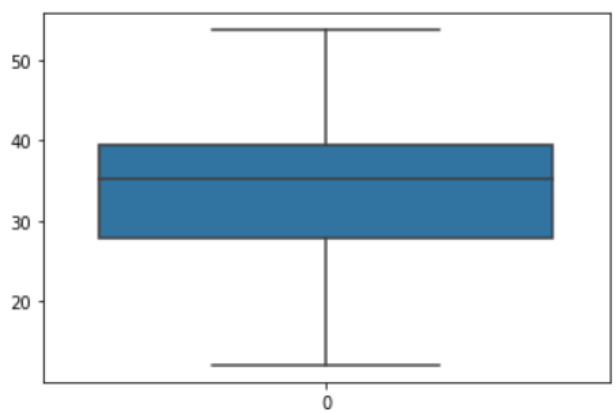
In [3]:
```
sns.kdeplot(car["MPG"])
plt.show()
```
executed in 268ms, finished 10:22:47 2021-08-30



In [4]:
```
car.MPG.describe()
```
executed in 22ms, finished 10:23:00 2021-08-30

Out[4]:
```
count    81.000000
mean     34.422076
std       9.131445
min      12.101263
25%      27.856252
50%      35.152727
75%      39.531633
max      53.700681
Name: MPG, dtype: float64
```

In [5]:
```
sns.boxplot(data=car["MPG"])
plt.show()
```
executed in 111ms, finished 10:23:23 2021-08-30



Inference;

- In boxplot Q2 is not accurate center, whisker is less negative side.
- Median (Q2) is nearer to Median(Q3), but not equal to it.

- Bell curve slightly skewed towards negative.
- MPG of Cars can follow normal distribution approximately (as mean and median are approximately same)

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

ANS;

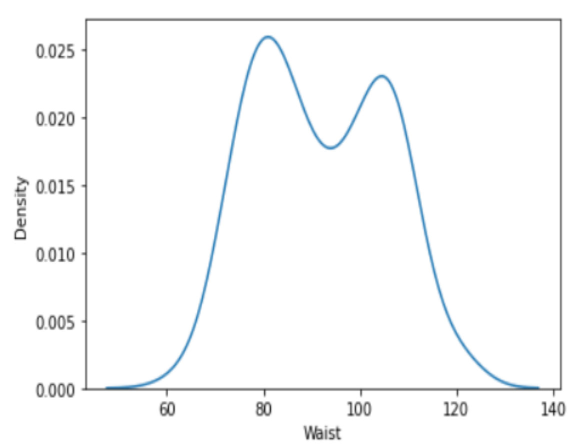
In [1]:
```python
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```
executed in 1.32s, finished 10:25:31 2021-08-30

In [2]:
```python
wc=pd.read_csv("wc-at.csv")
wc.head()
```
executed in 53ms, finished 10:25:41 2021-08-30

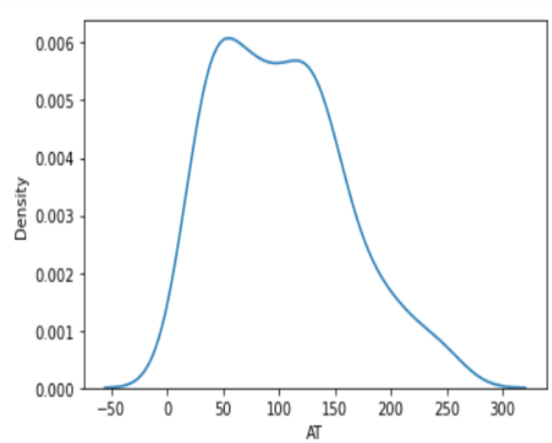
Out[2]:

	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

```python
sns.kdeplot(wc["Waist"])
plt.show()
```
executed in 207ms, finished 10:25:46 2021-08-30




```python
sns.kdeplot(wc["AT"])
plt.show()
```
executed in 160ms, finished 10:25:50 2021-08-30



In [5]:
```
wc.describe()
```
executed in 38ms, finished 10:25:52 2021-08-30

Out[5]:

	Waist	AT
count	109.000000	109.000000
mean	91.901835	101.894037
std	13.559116	57.294763
min	63.500000	11.440000
25%	80.000000	50.880000
50%	90.800000	96.540000
75%	104.000000	137.000000
max	121.000000	253.000000

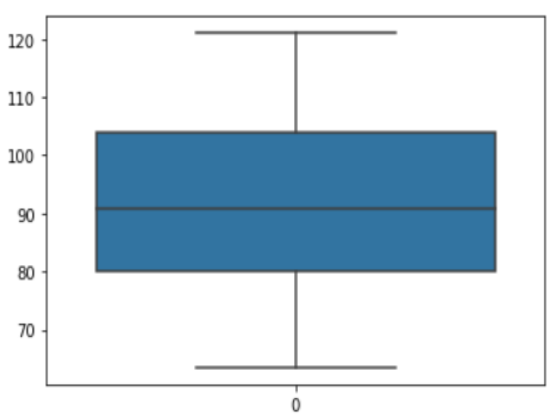
In [6]:
```
wc.Waist.skew(),wc.Waist.kurt()
```
executed in 46ms, finished 10:26:09 2021-08-30

Out[6]: (0.1340560824786468, -1.1026666011768886)

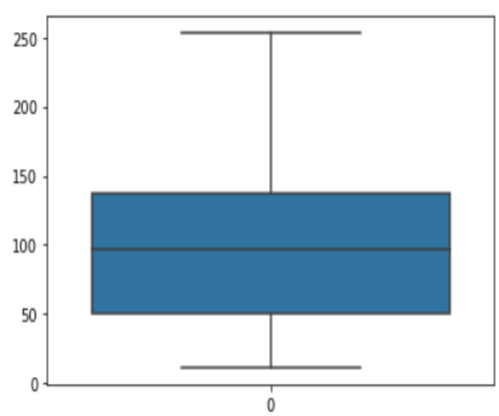
In [7]:
```
wc.AT.skew(),wc.AT.kurt()
```
executed in 29ms, finished 10:26:10 2021-08-30

Out[7]: (0.584869324127853, -0.2855757675045844 25)

```
sns.boxplot(data=wc["Waist"])
plt.show()
```
executed in 131ms, finished 11:11:07 2021-08-30




```
sns.boxplot(data=wc["AT"])
plt.show()
```
executed in 125ms, finished 11:11:11 2021-08-30



Inference;

- Both the (AT) and (Waist) data set are approximately equal to each other.
- mean > median (Right skewed)
- slightly positively skewed data.

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

ANS;

In [1]:
```
import pandas as pd
import numpy as np
from scipy import stats
```
executed in 925ms, finished 10:28:28 2021-08-30

In [2]:
```
stats.norm.ppf(0.95)
```
executed in 38ms, finished 10:28:28 2021-08-30

Out[2]:
```
1.6448536269514722
```

In [3]:
```
stats.norm.ppf(0.97)
```
executed in 16ms, finished 10:28:28 2021-08-30

Out[3]:
```
1.8807936081512509
```

In [4]:
```
stats.norm.ppf(0.8)
```
executed in 16ms, finished 10:28:28 2021-08-30

Out[4]:
```
0.8416212335729143
```

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

ANS;

In [1]:
```
import pandas as pd
import numpy as np
from scipy import stats
```
executed in 747ms, finished 10:29:36 2021-08-30

In [2]:
```
stats.t.ppf(0.975,24)
```
executed in 36ms, finished 10:29:36 2021-08-30

Out[2]:
```
2.0638985616280205
```

In [3]:
```
stats.t.ppf(0.98,24)
```
executed in 13ms, finished 10:29:36 2021-08-30

Out[3]:
```
2.1715446760080677
```

In [4]:
```
stats.t.ppf(0.995,24)
```
executed in 31ms, finished 10:29:37 2021-08-30

Out[4]:
```
2.796939504772804
```

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 buslbs 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;

Average light bulb (μ)=270

Sample bulb (n)=18

Average Sample (x)=260

Standard deviation (S)=90

T = (X – μ) / [s/√(n)]

T = (260-270)/ (90/18**0.5)

T = - 0.4714

Pt= - 0.4714, df=17

```
In [1]: import pandas as pd
        import numpy as np
        from scipy import stats
```

executed in 747ms, finished 10:29:36 2021-08-30

```
In [2]: p_value=stats.t.sf(abs(-0.4714),df=17)
        p_value
```

executed in 33ms, finished 10:34:46 2021-08-30

```
Out[2]: 0.32167411684460556
```

Probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 32.17%.