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)	Nominal Data

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	Nominal
Level of Agreement	Ordinal
IQ(Intelligence Scale)	Interval
Sales Figures	Ratio
Blood Group	Nominal
Time Of Day	Ordinal
Time on a Clock with Hands	Interval
Number of Children	Ratio
Religious Preference	Nominal

Barometer Pressure	Ratio
SAT Scores	Interval
Years of Education	Ratio

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

Ans. Probability = 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. Probability=

a) 1/36
b) 1/6
c) 5/8

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= 10/21

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 of candies for a randomly selected child

= 1 * 0.015 + 4*0.20 + 3 *0.65 + 5*0.005 + 6 *0.01 + 2 * 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.09

Expected number of candies for a 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.

Ans.

	Point	Score	Weight
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
S.D	0.534679	0.978457	1.786943
Range	2.17	3.911	8.4

Use Q7.csv file

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 = ∑ (Probability * Value)

∑ P(x).E(x)

There are 9 patients

Probability of selecting each patient = 1/9

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

= (1/9) (1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

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

Cars speed and distance

Use Q9_a.csv

Ans.

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

data=pd.read_csv("/content/Q9_a.csv")
```

```
print(data)

print(data.skew())

print(data.kurt())
```

```
    Index  speed  dist
0       1      4     2
1       2      4    10
2       3      7     4
3       4      7    22
4       5      8    16
5       6      9    10
6       7     10    18
7       8     10    26
8       9     10    34
9      10     11    17
10     11     11    28
11     12     12    14
12     13     12    20
13     14     12    24
14     15     12    28
15     16     13    26
16     17     13    34
17     18     13    34
18     19     13    46
19     20     14    26
20     21     14    36
21     22     14    60
22     23     14    80
23     24     15    20
24     25     15    26
25     26     15    54
26     27     16    32
27     28     16    40
28     29     17    32
29     30     17    40
30     31     17    50
31     32     18    42
32     33     18    56
```

```
33      34      18     76
34      35      18     84
35      36      19     36
36      37      19     46
37      38      19     68
38      39      20     32
39      40      20     48
40      41      20     52
41      42      20     56
42      43      20     64
43      44      22     66
44      45      23     54
45      46      24     70
46      47      24     92
47      48      24     93
48      49      24    120
49      50      25     85

Index    0.000000
speed   -0.117510
dist     0.806895
dtype: float64

Index   -1.200000
speed   -0.508994
dist     0.405053
dtype: float64
```

SP and Weight(WT)

Use Q9_b.csv

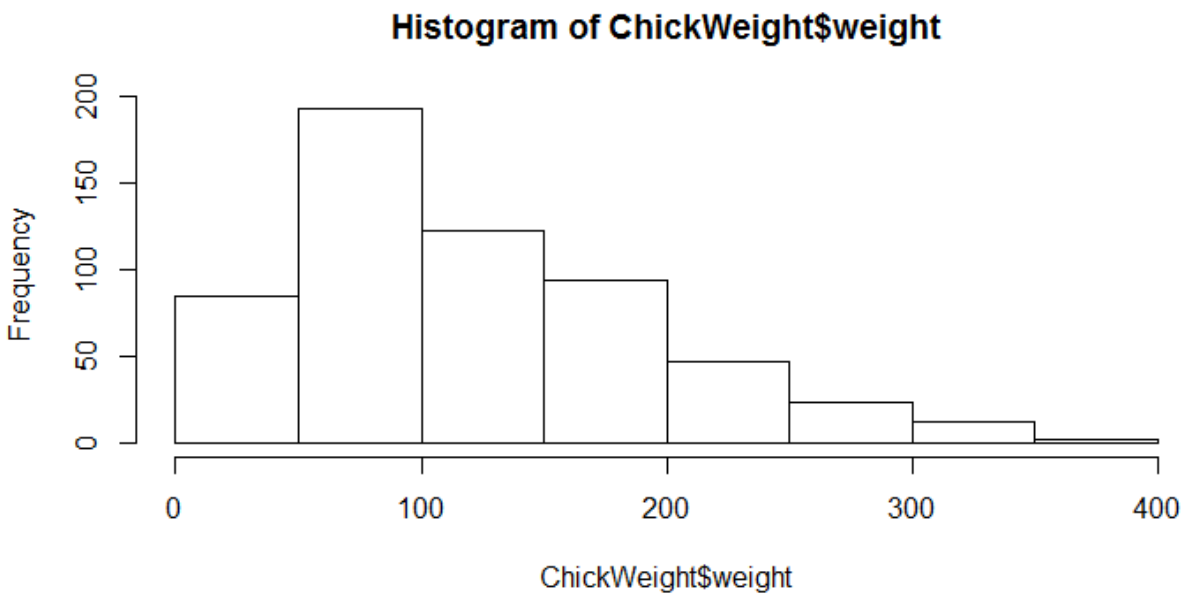
Ans.

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

data=pd.read_csv("/content/Q9_b.csv")
```

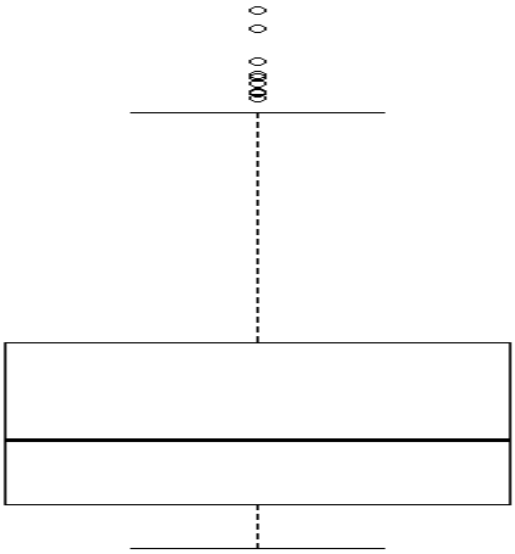
```
print(data)

print(data.skew())

print(data.kurt())
```

```
 Unnamed: 0         SP         WT
0           1  104.185353  28.762059
1           2  105.461264  30.466833
2           3  105.461264  30.193597
3           4  113.461264  30.632114
4           5  104.461264  29.889149
..        ...         ...        ...
76         77  169.598513  16.132947
77         78  150.576579  37.923113
78         79  151.598513  15.769625
79         80  167.944460  39.423099
80         81  139.840817  34.948615

[81 rows x 3 columns]
Unnamed: 0    0.000000
SP            1.611450
WT           -0.614753
dtype: float64

Unnamed: 0   -1.200000
SP            2.977329
WT            0.950291
dtype: float64
```

Q10) Draw inferences about the following boxplot & histogram



Ans. The histograms peak has right skewed and tail is on right. Mean > Median. We have outliers on the higher side.



Ans. The boxplot has outliers on the maximum side.

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.

```
import numpy as np
import pandas as pd
from scipy import stats
from scipy.stats import norm

# Average weight of Adults in Mexico with 94%
Confidense Interval
print(stats.norm.interval(0.94,200,30/(2000**0.5)))
```

(198.738325292158, 201.261674707842)

```
# Average weight of Adults in Mexico with 98%
Confidense Interval
print(stats.norm.interval(0.98,200,30/(2000**0.5)))
```

(198.43943840429978, 201.56056159570022)

```
# Average weight of Adults in Mexico with 96%
Confidense Interval
stats.norm.interval(0.96,200,30/(2000**0.5))
```

(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.

2) What can we say about the student marks?

Ans.

```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
import statistics

marks=[34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,
49,56]

print(statistics.mean(marks))

print(statistics.median(marks))

print(statistics.mode(marks))

print(statistics.stdev(marks))

print(statistics.variance(marks))

plt.hist(marks)
plt.grid()
plt.show()

plt.boxplot(marks)
plt.grid()
plt.show()
```

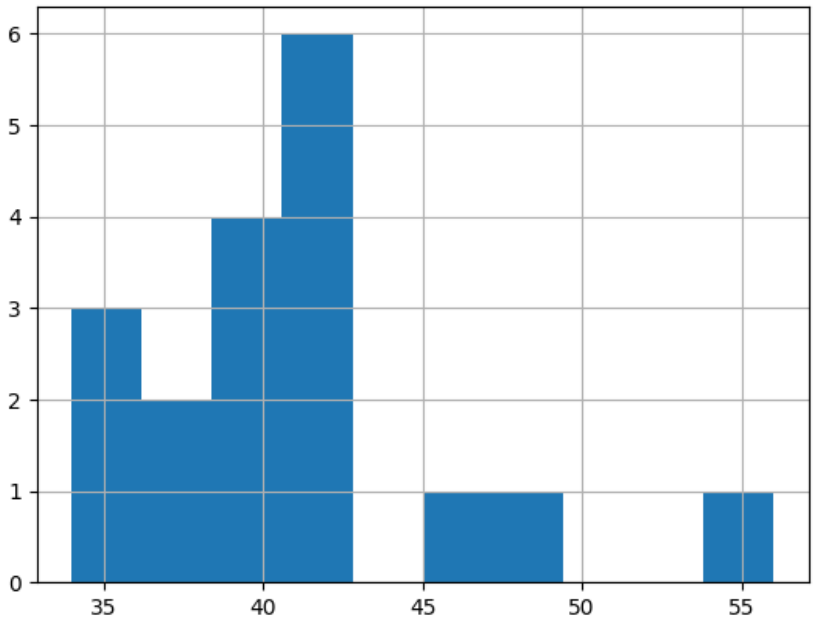
Mean= 41

Median= 40.5

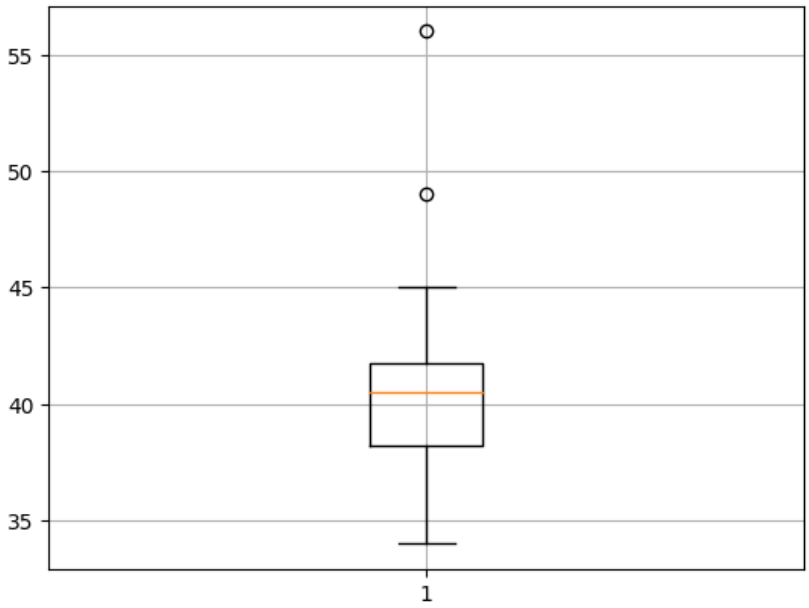
Variance= 25.529411764705884

Standard Deviation= 5.05266382858645

Histogram:-



Boxplot:-




```
#From above plot we can say that mean of marks of
student is 41 which is slightly greater than median.

#Most of the students got marks in between 41-42, there
are two outlier 49, 56.
```

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

Ans. If the mean and median of dataset are equal, it indicates that distribution is symmetrical and has zero skewness.

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

Ans. When the mean is greater than the median, it indicates a right-skewed distribution with a tail extending towards higher values.

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

Ans. When the median is greater than the mean, it indicates a left-skewed distribution with a tail extending towards lower values.

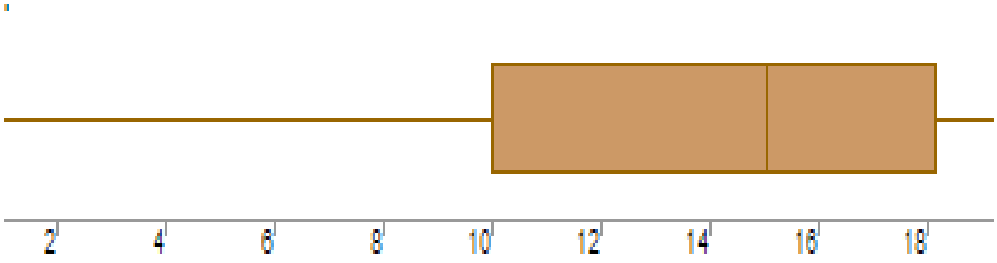
Q16) What does positive kurtosis value indicates for a data?

Ans. A positive kurtosis value indicates that a data set has heavier tails or outliers compared to a normal distribution, indicating a distribution with more extreme values or observations.

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

Ans. A negative kurtosis value indicates that a data set has lighter tails or fewer outliers compared to a normal distribution, indicating a distribution with fewer extreme values or observations**.**

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



What can we say about the distribution of the data?

Ans. The above Boxplot is not normally distributed the median is towards the higher value.

What is nature of skewness of the data?

Ans. The data is a skewed towards left. The whisker range of minimum value is greater than maximum.

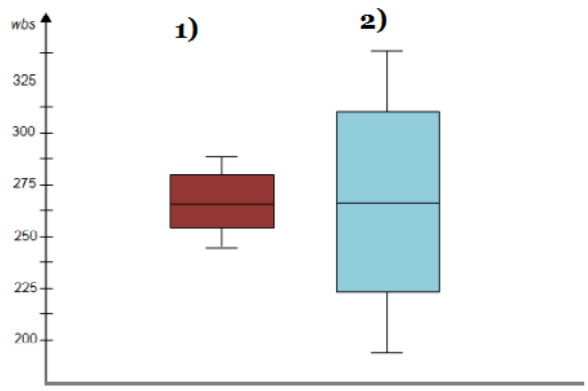
What will be the IQR of the data (approximately)?

Ans. The Inter Quantile Range = Q3 Upper quartile – Q1 Lower Quartile

= 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. 1) There are no outliers.

2) Both the box plot shares the same median that is approximately in a range between 275 to 250 and they are normally distributed with zero to no skewness neither at the minimum or maximum whisker range.

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.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
```

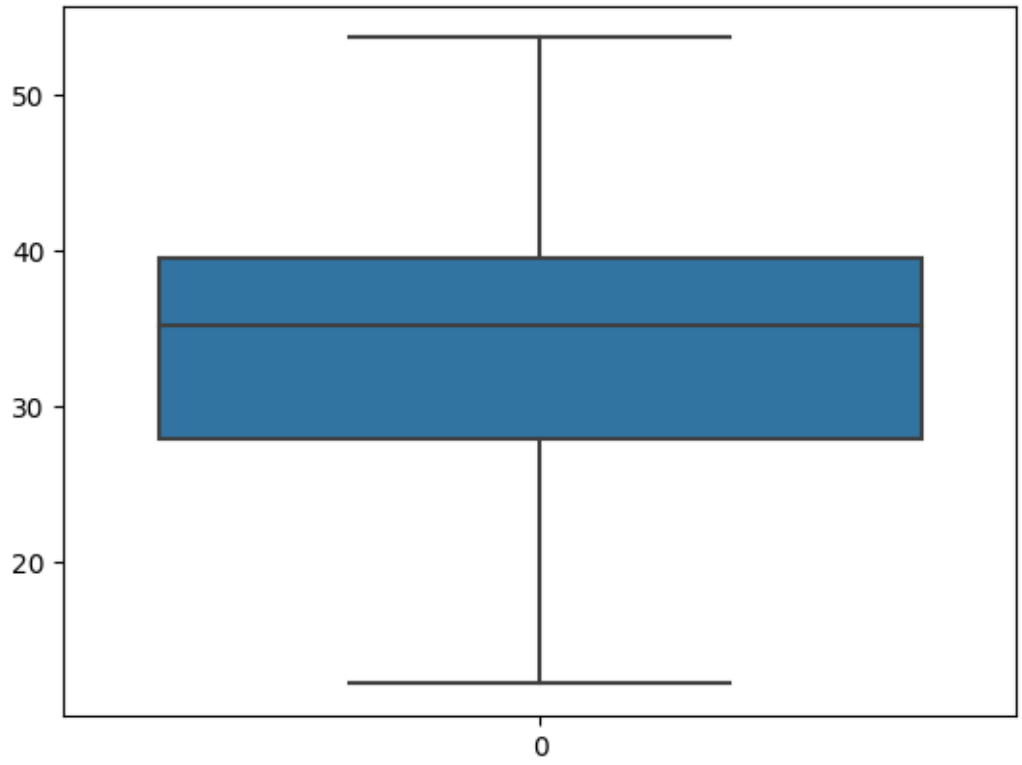
```
from scipy.stats import norm
data=pd.read_csv("/content/Cars.csv")

data

sns.boxplot(data.MPG)

# P(MPG>38)
print(1-
stats.norm.cdf(38,data.MPG.mean(),data.MPG.std()))

# P(MPG<40)
print(stats.norm.cdf(40,data.MPG.mean(),data.MPG.std())
)

# P (20<MPG<50)
print(stats.norm.cdf(0.50,data.MPG.mean(),data.MPG.std(
))-stats.norm.cdf(0.20,data.MPG.mean(),data.MPG.std()))
```

	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 x 5 columns]

```
0.34759392515827137

0.7293498762151609
```

```
1.2430968797327491e-05
```



Q 21) Check whether the data follows normal distribution

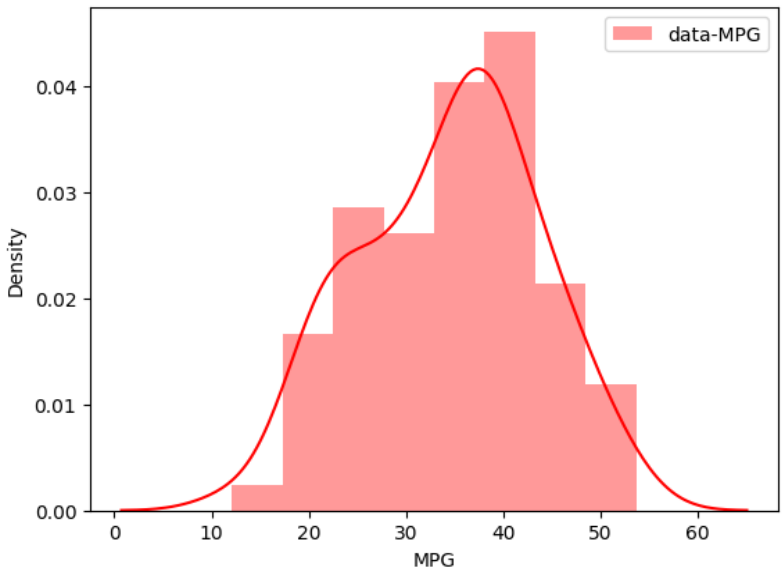
a) Check whether the MPG of Cars follows Normal Distribution
Dataset: Cars.csv

Ans.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import norm

data=pd.read_csv("/content/Cars.csv")

print(data)
```

```
sns.distplot(data.MPG, label='data-MPG',color="red")
plt.xlabel('MPG')
plt.ylabel('Density')
plt.legend();

print(data.MPG.mean())

print(data.MPG.median())
```

	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



Mean=34.42207572802469

Median=35.15272697

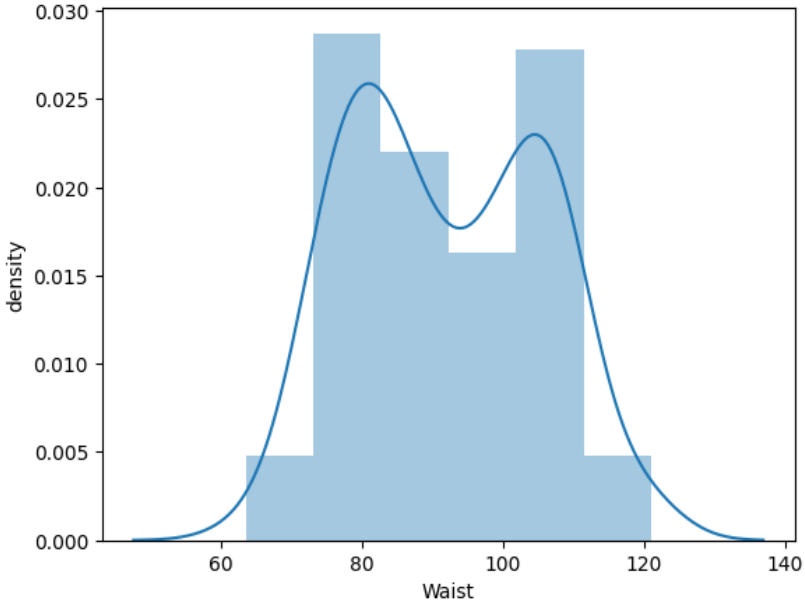
b) Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution
Dataset: wc-at.csv

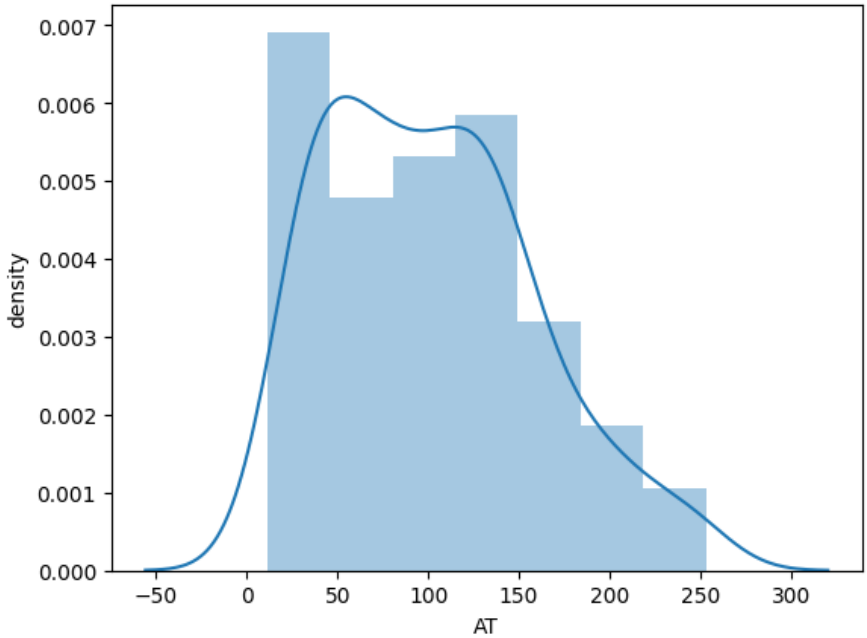
Ans.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

data=pd.read_csv("/content/wc-at.csv")
print(data)


# Plotting distribution for Waist Circumference (Waist)
print(sns.distplot(data.Waist))
print(plt.ylabel('density'));

# Plotting distribution for Adipose Tissue (AT)
sns.distplot(data.AT)
plt.ylabel('density');


print(data.Waist.mean() , data.Waist.median())

print(data.AT.mean() , data.AT.median())
```

```
      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 x 2 columns]
```






```
(91.90183486238531, 90.8)
```

```
(101.89403669724771, 96.54)
```

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

Ans.

```
from scipy import stats
from scipy.stats import norm
```

```
# Z-score of 90% confidence interval
stats.norm.ppf(0.95)
```

```
1.6448536269514722
```

```
# Z-score of 94% confidence interval
stats.norm.ppf(0.97)
```

```
1.8807936081512509
```

```
# Z-score of 60% confidence interval
stats.norm.ppf(0.8)
```

```
0.8416212335729143
```

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

Ans.

```
from scipy import stats
from scipy.stats import norm
```

```
# t scores of 95% confidence interval for sample size
of 25
stats.t.ppf(0.975,24)
```

```
2.0638985616280205


# t scores of 96% confidence interval for sample size
of 25
stats.t.ppf(0.98,24)

2.1715446760080677

# t scores of 99% confidence interval for sample size
of 25
stats.t.ppf(0.995,24)


2.796939504772804
```

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.

```
from scipy import stats
from scipy.stats import norm


# Assume Null Hypothesis is: Ho = Avg life of Bulb >=
260 days
```

```
# Alternate Hypothesis is: Ha = Avg life of Bulb < 260
days


# find t-scores at x=260; t=(s_mean-
P_mean)/(s_SD/sqrt(n))
t=(260-270)/(90/18**0.5)
t

0.4714045207910317

# Find P(X>=260) for null hypothesis


# p_value=1-stats.t.cdf(abs(t_scores),df=n-1)... Using
cdf function
p_value=1-stats.t.cdf(abs(-0.4714),df=17)
p_value

0.32167411684460556

#  OR p_value=stats.t.sf(abs(t_score),df=n-1)... Using
sf function
p_value=stats.t.sf(abs(-0.4714),df=17)
p_value


0.32167411684460556
```