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	Continuous
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	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.

Possible Outcomes = (HHH, HHT, HTH, THH, TTH, THT, HTT, TTT)

Desired Outcomes = (HHT, HTH, THH)

Probability of getting two heads and one tail = $\frac{3}{8}$ = $\underline{0.375}$

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.

a) Sum being one in the given case is not possible. Hence

   Probability that the sum is equal to 1 = **0**

b) {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}

   Probability that the sum is less than or equal to four is

$$= \frac{6}{6\times6} = \frac{1}{6} = \underline{0.167}$$

c) {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}

   Probability that the sum is divisible by 2 and 3

$$= \frac{6}{36} = \underline{0.167}$$

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 number of ways of drawing two balls at random = 7C2

Total number of ways of drawing two balls that are not blue = 5C2

Probability that none of the balls drawn is blue when two balls are drawn at random $= \frac{5C2}{7C2} = \frac{10}{21} = \underline{0.4760}$

(OR)

Probability that none of the balls drawn is blue when two balls are drawn at random $= \frac{5}{7} \times \frac{4}{6} = \frac{10}{21} = \underline{0.4760}$

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.

CHILD	Expected Value
A	$1 \times 0.015 = 0.015$
B	$4 \times 0.20 = 0.80$
C	$3 \times 0.65 = 1.95$
D	$5 \times 0.005 = 0.025$
E	$6 \times 0.01 = 0.06$
F	$2 \times 0.120 = 0.240$
	$\Sigma = 3.09$

The Expected number of candies for a randomly selected child = $\underline{3.09} \approx \underline{3}$

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.

Parameter	Points	Score	Weigh
Mean	3.596563	3.217250	17.848750
Median	3.695	3.325	17.710
Mode	3.92	3.44	17.02
Standard Deviation	0.534679	0.978457	1.786943
Variance	0.285881	0.957379	3.193166
Range	2.170	3.911	8.4

- The given dataset comprises a Car Collection with 31 entries and their ratings. Points of Cars ranges from a maximum point of 4.93 (Honda Civic) and a minimum of 2.76 (Dodge Challenger) with an average of 3.6 Points.
- The Scores have an average of about 3.21 and range between 5.42 (Lincoln Continental) and 1.51(Lotus Europa).
- Weights of the cars exhibits much more variation over values from 22.9 (Merc 230) to 14.5 (Ford Panthera L) with an average weight of 17.84.
- Mean and median are closer to each other indicating an absence of extreme values in the data set.

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 = 145.333333

```
In [5]:
1 import pandas as pd
2 import numpy as np

In [2]:
1 Patient_list = [108, 110, 123, 134, 135, 145, 167, 187, 199]

In [3]:
1 Q8 = pd.DataFrame(Patient_list)
2 Q8
```

Out[3]:

	0
0	108
1	110
2	123
3	134
4	135
5	145
6	167
7	187
8	199

```
In [7]:
1 expected_value = np.mean(Q8)
2 expected_value

Out[7]: 0    145.333333
        dtype: float64
```

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

a. **Car's speed and distance**

   Use Q9_a.csv

b. **SP and Weight (WT)**

   Use Q9_b.csv

Ans.

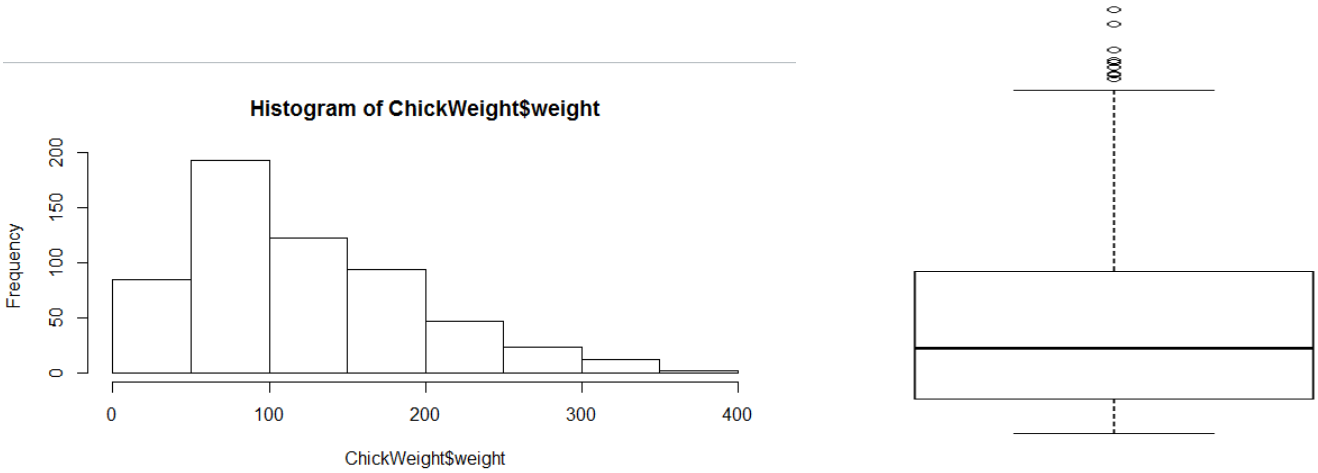
	Speed	Distance
Skewness	-0.11395477	0.78248352
Kurtosis	-0.57714742	0.24801866

- Negative value of Skewness for speed implies that it is left tailed with most of the data towards the right. Also, the negative value for Kurtosis reveals that the data has a wider peak.
- Distance has positive values for both Skewness and Kurtosis which implies that it is right tailed with extreme values to the right. Positive kurtosis reveals the narrower peak in the distribution.

	SP	WT
Skewness	1.58145368	-0.60330993
Kurtosis	2.72352149	0.81946588

- SP has high positive values for both Skewness and Kurtosis which means that it is right tailed with most values towards the left. High Kurtosis value is pointing towards the narrow peak and the thinner tails.
- WT has a negative Skewness value and a positive Kurtosis value which means that it is left tailed and hence a concentration in values to the right. Positive kurtosis can be used to predict a comparatively narrow peak.

Q10) Draw inferences about the following boxplot & histogram



Ans.

The Histogram :

- The histogram portrays a rather asymmetric distribution between Chick Weight and their frequency of occurrence with a concentration of data leftward and a right tail. These can be used to infer a positive value of Skewness.
- The frequency of data is higher between the 50 – 100 interval where mean is likely situated. Frequency of occurrence for higher values of Chick Weight is lower.
- The data seems to have a wider peak that could be pointing at a lower Kurtosis value.

The Boxplot :

- The boxplot captures an asymmetric distribution that can be observed hanging closer to the lower limit of the graph.
- The lower whisker seems much shorter than it's upper counterpart. This means that the upper quartile is longer than the lower quartile.
- Few extreme values (Outliers) can be observed at the upper limit of the graph outside the boxplot. The median is situated closely towards the lower extreme of the graph**.**

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 numpy as np
         2 import scipy.stats as st
```

```
In [5]:  1 [round(x,2) for x in st.norm.interval(alpha=0.94, loc=200, scale=30)]
Out[5]: [143.58, 256.42]

In [6]:  1 [round(x,2) for x in st.norm.interval(alpha=0.98, loc=200, scale=30)]
Out[6]: [130.21, 269.79]

In [7]:  1 [round(x,2) for x in st.norm.interval(alpha=0.96, loc=200, scale=30)]
Out[7]: [138.39, 261.61]
```

C.I.		Interval
94%	→	143.58, 256.42
96%		130.21, 269.79
98%		138.39, 261.61

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?**

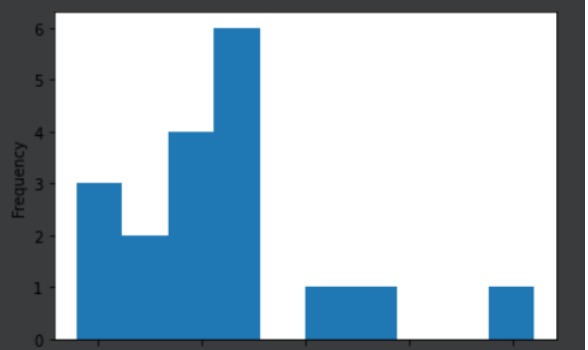
Ans.

1) Mean = 41
   Median = 40.50
   Variance = 25.529
   Standard Deviation = 5.053
2) From the histogram, it can be observed that the highest frequency of occurrence is between the 40 to 43 interval. Also, it can be noted from the data that the central tendencies mean and median occurs in this very same interval. The highest marks scored was 56 and the lowest being 34. The distribution of the scores is rather asymmetric in nature.

```
In [3]: 1 Qn12.describe()
Out[3]:
```

	0
count	18.000000
mean	41.000000
std	5.052664
min	34.000000
25%	38.250000
50%	40.500000
75%	41.750000
max	56.000000

```
In [7]: 1 plt.hist(Qn12)
        2 plt.xlabel('Scores')
        3 plt.ylabel('Frequency')
Out[7]: Text(0, 0.5, 'Frequency')
```




```
In [4]: 1 Qn12.median()
Out[4]: 0    40.5
        dtype: float64
```

```
In [9]: 1 round(Qn12.var(), 3)
Out[9]: 0    25.529
        dtype: float64
```

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

Ans) In the given case, data is symmetric and normally distributed. Hence the Skewness will be zero.

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

Ans) Mean > Median implies that the data is concentrated towards the right and thus having a left tail. Therefore, the Skewness will be negative**.**

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

Ans) Median > Mean implies that the data is concentrated towards the left and thus having a right tail. Therefore, the Skewness will be positive**.**

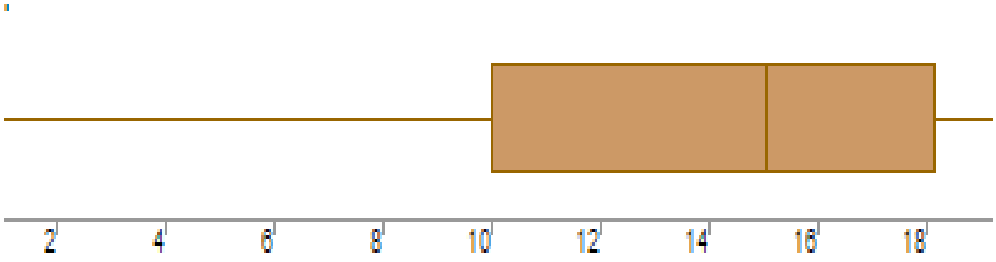
Q16) What does positive kurtosis value indicates for a data?

Ans) Positive Kurtosis (leptokurtic) means that the distribution has flatter tails and thin sharp peaks**.**

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

Ans) Negative Kurtosis (platykurtic) means that the distribution has thinner tails and wider flat peaks**.**

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

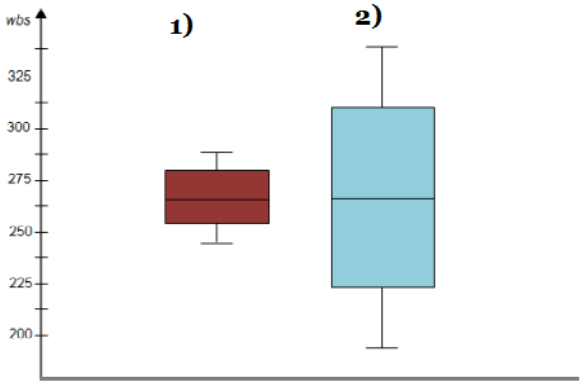


i. **What can we say about the distribution of the data?**
ii. **What is nature of skewness of the data?**
iii. **What will be the IQR of the data (approximately)?**

Ans.

i. The distribution of the data seems asymmetric and is hanging closer to the lower extreme. It can be inferred that the data is NOT normally distributed.
ii. Here the data is concentrated toward the lower end and hence right tailed. Therefore, the distribution has Positive Skewness.
iii. From the boxplot, it can be observed that the Inner Quartile Range is approximately 10 to 18.

Q19) Comment on the below Boxplot visualizations?



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

Ans.

- Both of the Boxplots (1) and (2) seems to be symmetric and normally distributed.
- Median of the boxplots (1) and (2) is almost the same.
- The range of Boxplot (1) is much bigger compared to Boxplot (1)
- The IQR of Boxplot (1) is around 250 to 280 (roughly) and that of Boxplot (2) is around 220 to 310 (again, roughly).

Q20) 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.

- The first step is to find the Mean and Standard deviation of column 'MPG'

```
In [14]:  1 round(1 - ss.norm.cdf(38, 34.422076,9.131445), 4)
Out[14]: 0.3476

In [19]:  1 round(ss.norm.cdf(40, 34.422076,9.131445), 4)
Out[19]: 0.7293

In [18]:  1 round(ss.norm.cdf(50, 34.422076,9.131445) - (1 - ss.norm.cdf(20, 34.422076,9.131445)), 4)
Out[18]: 0.0131
```

- **P(MPG>38)** = 0.3476
- **P(MPG<40)** = 0.7293
- **P(20<MPG<50)** = 0.0131

Q21) Check whether the data follows normal distribution

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

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

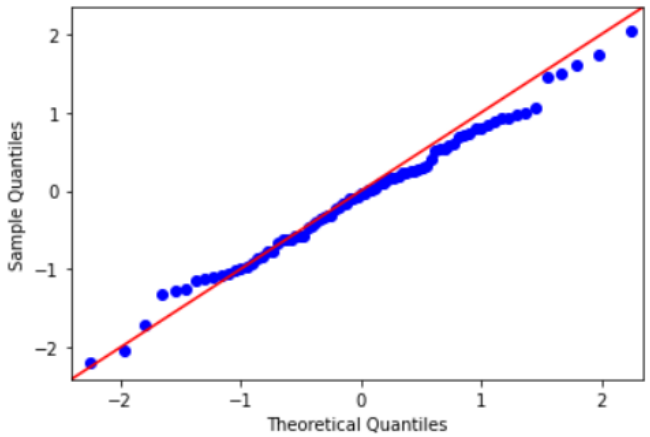
Ans.

a)

- To check for the normality of the distribution, we can use Q-Q (Quantile Quantile) plot. If our data comes from a normal distribution, we should see all the points sitting on the straight line.
- We can also use Shapiro-Wilk test to check the normality of a distribution
  - If the p-value ≤ 0.05, then we can assume the distribution of our variable is not normal/gaussian.
  - If the p-value > 0.05, then we assume the distribution of our variable is normal/gaussian.

- Plotting the QQ Plot for MPG. We get,

```
In [3]:  1 Qn21.MPG = norm.rvs(size=81)
         2 sm.qqplot(Qn21.MPG, line='45')
         3 pylab.show()
```



✓ Since the data follows a straight line (approximately), the dataset might be normally distributed.

Running Shapiro test on MPG, we get the result. We get,

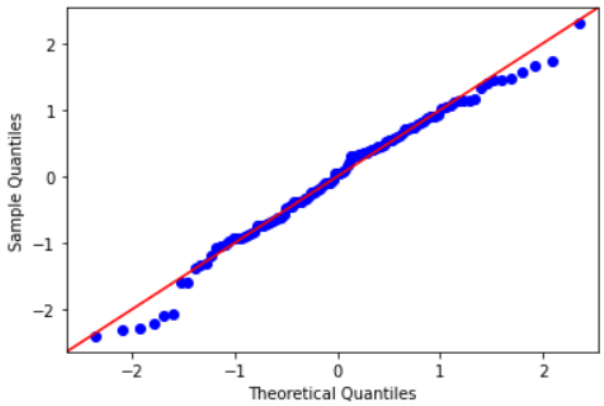
```
ShapiroResult(statistic=0.9910812973976135,
pvalue=0.853973925113678)
```

- ✓ Since p>0.05, MPG probably follows normal distribution.

b)

- Plotting the QQ Plot for MPG. We get,

```
In [7]:
1 Qn21a.AT = norm.rvs(size=109)
2 sm.qqplot(Qn21a.AT, line = '45')
3 pylab.show()
```



- ✓ Since the data follows a straight line (approximately), the dataset might be normally distributed.

- **R**unning Shapiro test on AT. The result is,

```
ShapiroResult(statistic=0.9863042831420898,
pvalue=0.3318818211555481)
```

- ✓ Since p>0.05, AT probably follows normal distribution.

Q22) Calculate the Z scores of 90% confidence interval, 94% confidence interval, 60% confidence interval

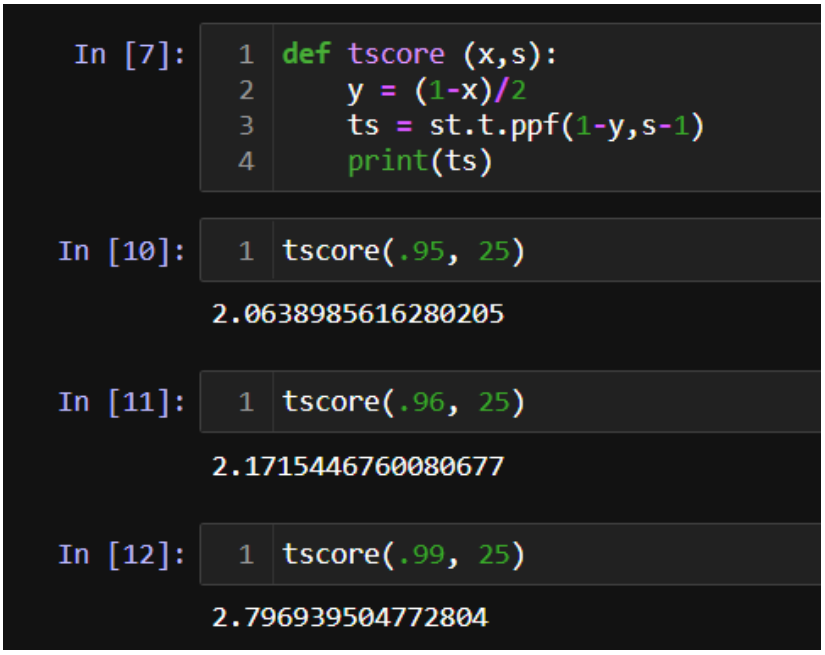
Ans.

```
In [2]:
1 def zscore(x):
2     y=(1-x)/2
3     s = st.norm.ppf(1-y)
4     print(s)

In [3]:
1 zscore(0.94)

1.8807936081512509

In [4]:
1 zscore(0.6)

0.8416212335729143

In [5]:
1 zscore(0.9)

1.6448536269514722
```

C.I.	Z Score
90%	1.644854
94%	1.880794
60%	0.841621

Q23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

Ans.



C.I.	T Score
95%	2.063898
96%	2.171545
99%	2.796939

Q24) 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.

Here, we have to find the probability that the sample bulbs have a life < 260

The method is to find the t score and the corresponding p value or the probability.

$$\text{T- Score} = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}}$$

Substituting all the values,

$$\text{T- Score} = \frac{260 - 270}{\frac{90}{\sqrt{18}}}$$

$$\text{T Score} = \underline{0.4714045208}$$

Corresponding to the above T-Score and the degrees of freedom (df=18-1=17), the p value can be found.

```
In [3]:  1  ss.t.sf(abs(-0.4714045207910317), df=17)

Out[3]: 0.32167253567098364
```

$$\text{P} = 0.\underline{32167253567098364}$$

So, The probability that 18 randomly selected bulbs would have an average life of no more than 260 days is approximately <u>32.17%</u>