Q1) Identify the Data type for the Following:

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

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	Interval
Blood Group	Nominal
Time Of Day	Interval
Time on a Clock with Hands	Ratio
Number of Children	Interval

Religious Preference	Nominal
Barometer Pressure	Interval (Since pressure is zero in outer space – perfect vacuum)
SAT Scores	Interval
Years of Education	Interval

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

Ans: Sample space for 3 coin tosses. (H- heads, T -tails)

HHH
HHT
HTH
HTT
THH
THT
TTH
TTT

No of times 2 heads and one tail occur = 3

Tot no of events = 8

The probability of getting 2 heads and 1 tail is:

$$\mathrm{P}(2\mathrm{H}, 1\mathrm{T}) = \frac{3}{8} = \mathbf{0.375}$$

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: Sample space for rolling two dice

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

a) Probability that the sum = 1 =

$$\mathrm{P}(\text{sum} = 1) = \frac{0}{36} = \mathbf{0}$$

b) Less than or equal to 4

$$\mathrm{P}(\text{sum} \leq 4) = \frac{6}{36} \sim 0.167 \; or \sim \mathbf{16.7}\%$$

c) Sum is divisible by 2 and 3

$$\mathrm{P}(\text{sum div by 2\&3}) = \frac{6}{36} \sim 0.167 \; or \sim \mathbf{16.7}\%$$

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: Possible ways of not picking blue balls are:

- 2 red balls.
- 2 green balls out of 3.
- 1 red ball out of 2 and 1 green ball out of 3.

Total no of ways of picking 2 balls = $\binom{7}{2} = 21$

$$\mathrm{P}(no\ blue) = \frac{\binom{2}{2} + \binom{3}{2} + \left\{\binom{2}{1} \times \binom{3}{1}\right\}}{\binom{7}{2}} = \frac{10}{21} \sim 0.476 \; or \; \mathbf{47.6}\%$$

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:

Let x be the number of candies for a randomly selected child.

CHILD	Candies count (xi)	Probability(P(xi)	xi*P(xi)
A	1	0.015	0.015
B	4	0.20	0.8
C	3	0.65	1.95
D	5	0.005	0.025
E	6	0.01	0.06
F	2	0.120	0.24
	Tot no of candies N = **Σ** xi = 21	**Σ** xi*P(xi)	=3.09

The expectation value of x is given by.

$$\langle x \rangle = \sum_i x_i P(x_i)$$

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.

Use Q7.csv file

Ans:

Feature	Mean	Median	Mode	Variance	StDev	min	max	Range
Points	3.597	3.695	3.92	0.2859	0.5347	2.76	4.93	2.17
Score	3.217	3.325	3.44	0.9574	0.9785	1.513	5.424	3.911
Weigh	17.85	17.71	17.02	3.193	1.787	14.5	22.9	8.4

- The data set may be regarding how the cars fared in crash tests.

- Not much difference in mean, median, and mode for all three features. This indicates a symmetric distribution about the mean.
- Distribution is narrow for all three since the standard deviation is much smaller compared to the range, which implies a smaller variability. Maybe most of the cars were manufactured with similar safety standards

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: Here the expected value is same as the average value. So,

$$\langle x \rangle = \frac{1}{N} \sum_i x_i$$

$\mathbf{\Sigma}$ xi = 1308

N = 9

Expected value, $\langle x \rangle \sim \mathbf{145.3}$

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

Cars speed and distance

Use Q9_a.csv

Ans. Note: The kurtosis values are "fisher's kurtosis) computed from pandas. This can also be seen as excess kurtosis.

Fisher's kurtosis = Pearson's kurtosis – 3

Feature	Skewness	Kurtosis
speed	-0.118	-0.509
dist	0.807	0.405

a) Speed:
   Distribution has a slight negative skewness which means it has a slight tail towards the left.
   The kurtosis is slightly negative which means the tails are slightly thinner than a normal distribution's tails.
   Due to the slight negative skewness and thinner tails, it may contain some low outliers which would be affecting the average speed calculations.

b) Distance:
   Distribution has a positive skewness which means it has a tail towards the right
   It also has a slight positive kurtosis which means it has fatter tails at the right than a normal distribution.
   Due to the positive skewness and the small positive kurtosis, there may be few high outliers.

SP and Weight(WT)

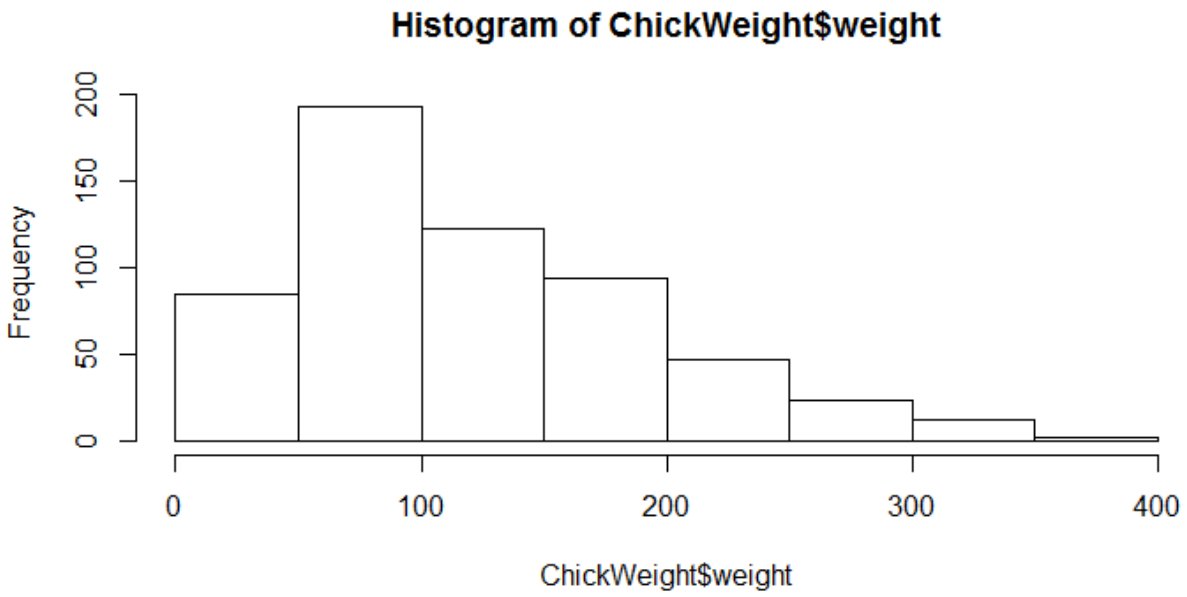
Use Q9_b.csv

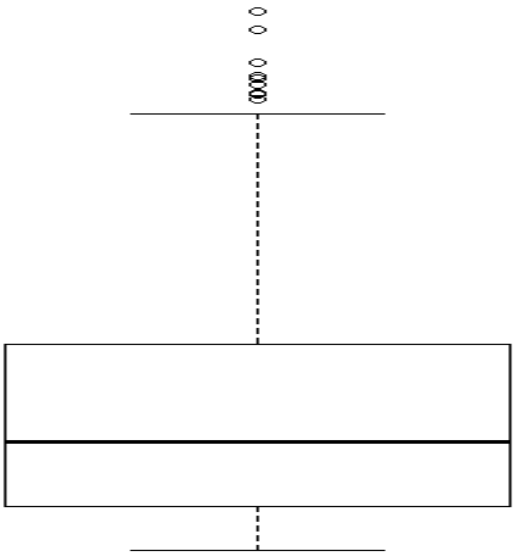
Ans.

Feature	**Skewness**	**Kurtosis**
SP	1.61	2.98
WT	-0.615	0.950

a) SP:
   The distribution has a fairly large positive skewness and a high kurtosis. This means that it is right tailed and the tails are fatter than normal distribution. We can expect a lot of high outliers in this distribution.
b) WT:
   The distribution has a negative skewness but positive kurtosis. This means that it is left skewed or left tailed but has fatter tails compared to the normal distribution. We can expect the distribution to have low outliers and a few high outliers as well.

Q10) Draw inferences about the following boxplot & histogram





Ans.

The data may have been collected from a poultry farm.

The median weight of the chickens can be around 80 or 90. By looking at the histogram and box plot we see that the distribution is left tailed or left skewed. There are high outliers but no low outliers. It seems like some chickens are over eating or gaining a lot of weight compared to others or the food seems to increase the weight of the chickens.

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. Given:

- sample size, n = 2000
- Population size, N =3000000
- Sample mean, $\overline{x}$ = 200Lbs
- Sample SD, s = 30Lbs

Population standard deviation is unknown but n>>30. Thus we can safely use normal distribution to obtain the confidence interval.

$$\text{Formula: } confidence\ interval = \overline{x} \pm \left(z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}\right)$$

Note: $z_{\frac{\alpha}{2}}$ is used since we want to find the confidence interval, thus have to account for both the tails or both halves of the distribution.

(1-α) (%)	α (%)	$\frac{\alpha}{2}$ (%)	$\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$	$\left(z_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}\right)$	Confidence interval (Lbs.)
94	6	3	0.03	± 1.88	1.3	198.7 - 201.3
96	4	2	0.02	± 2.05	1.4	198.6 - 201.4
98	2	1	0.01	± 2.30	1.5	198.5 - 201.5

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:

Mean	Median	mode	Variance	StDev
41	40.5	41	24.11	4.91

The given student marks are normally distributed since Mean median and mode are more or less equal.

Many students have scored around 41 and 42, thus if we pick a student at random, it is likely that his/her score will be this or +/- 1 standard deviation.

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

Ans. If the mean and median are same, the distribution is symmetrical and the skewness is **zero**.

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

Ans. When mean is greater than the median the distribution is **skewed to the right** or is **positively skewed**.

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

Ans. When mean is greater than the median the distribution is **skewed to the left** or is **negatively skewed**.

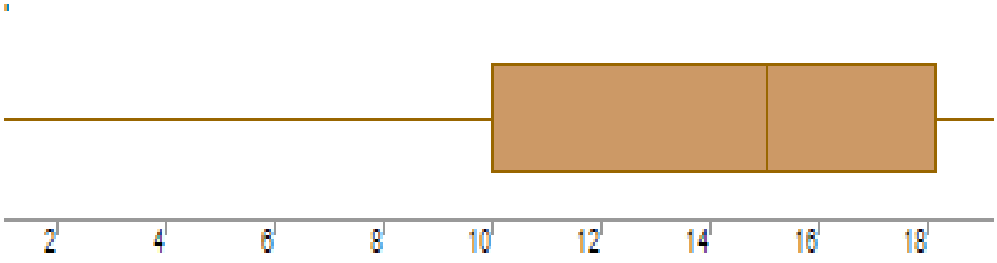
Q16) What does positive kurtosis value indicates for a data?

Ans. For a data, positive kurtosis means, the tail is much fatter than than the normal curve. Datasets with positive kurtosis tend to have outliers.

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

Ans. For a data, negative kurtosis means, the tail is much thinner than than the normal curve. Datasets with negative kurtosis lacks outsider.

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



What can we say about the distribution of the data?

Ans. The distribution is not symmetric(normal). The median is roughly 15. The range is much larger than the interquartile range. Since the whiskers are long, also the box, the distribution has a large spread.

What is nature of skewness of the data?

Ans. The distribution is left skewed or left tailed with low lying outliers.

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

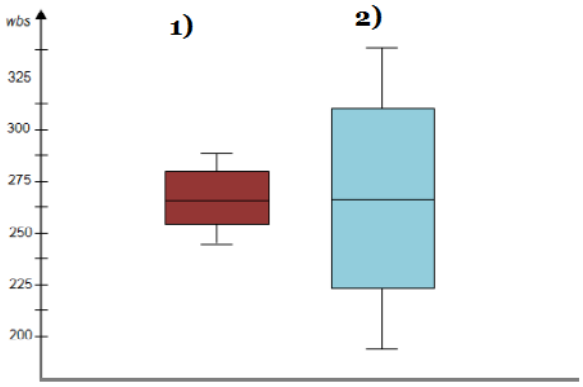
Ans. IQR = Q3 – Q1

Q1 = 10

Q3 = 18

IQR = 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.

- Both distributions have the same median but distribution 2 has a much larger spread than distribution 1. In other words, data points are much more densely clustered around the median for distribution 1 compared to distribution 2 as seen by comparing the IQR's i.e IQR of distribution 1 is lesser than IQR of distribution 2
- Both distributions are roughly symmetric. Distribution 1 may seem a little asymmetric since the rectangle above the median line has a slightly larger area, but we can approximate both by normal distributions and answer question related to probabilities and predictions.
- There are no outliers in both

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

Using pandas describe funvction, we can calculate the count, mean and standard deviation (sample StDev):

Count	**Mean**	**StDev**
81	34.422076	9.131445

a. P(MPG>38)
Ans: P(MPG>38) = 1 – P(MPG<38)
Now, P(MPG<38) can be found using scipy stats module
P(MPG<38) = 0.6524060595854699
1 – P(MPG<38) = 0.34759394041453007 ~ 0.348 or ~**34.8%**

```
from scipy import stats
p_less_38 = stats.norm.cdf(x=38, loc=34.422076,
scale=9.131445) # x -> val, loc -> mean, scale ->
stdev
p_grt_38 = 1-p_less_38
print(p_grt_val)
>> 0.34759394041453007
```

b. P(MPG<40)

Ans. P(MPG<40) = 0.7293498604157946 ~ 0.729 ~ **72.9%**

```
# b.P(MPG<40)
p_less_val = stats.norm.cdf(x=40, loc=34.422076,
scale=9.131445)
print(p_less_val)
>> 0.7293498604157946
```

c. P (20<MPG<50)

Ans. P (20<MPG<50) = P(MPG<20) - P(MPG<50)

P(MPG<20) = 0.05712377822429007

P(MPG<50) = 0.9559926858516099

P (20<MPG<50) = 0.8988689076273199 ~ 0.899 or ~ **89.9%**

```
# P (20<MPG<50)
p_less_20 = stats.norm.cdf(x=20, loc=34.422076,
scale=9.131445)
print(p_less_20)
>> 0.05712377822429007

p_less_50 = stats.norm.cdf(x=50, loc=34.422076,
scale=9.131445)
print(p_less_50)
>> 0.9559926858516099

p_bw_20_50 = p_less_50 - p_less_20
print(p_bw_20_50)
>> 0.8988689076273199
```

Q 21) Check whether the data follows normal distribution

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

Mean	**Median**	**Mode**	**Skewness**	**Kurtosis**
34.422076	35.152727	~38	-0.18	-0.61

Since, the mean median ad mode are close to each other and there is only a slight negative skewness**, we can approximate the 'MPG' distribution to a normal distribution**. The negative kurtosis indicates that the tail is a little thinner than normal, at the left.

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

Ans.

Mean	**Median**	**Mode**	**Skewness**	**Kurtosis**
101.894037	96.54	~ 43, ~123	0.58	-0.29

The mean is greater than the median in this distribution, which means the distribution is right tailed, and it has two modes. This is further validated by a positive skewness greater than 0.5. Thus, the '**AT' distribution is not normal**. A negative kurtosis indicates that the tail is a little thinner than the normal.

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

Ans.

(1-α) (%)	α (%)	$\frac{\alpha}{2}$ (%)	$\frac{\alpha}{2}$	$1-\frac{\alpha}{2}$	$z_{\frac{\alpha}{2}}$
60	40	20	0.2	0.80	± 0.84
90	10	5	0.05	0.95	± 1.64
94	6	3	0.03	0.97	± 1.88

Note: Used `stats.norm.ppf()`

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

Ans. Given n = 25, degrees of freedom, df = n-1 = 25-1 = 24.

(1-α) (%)	α (%)	$\frac{\alpha}{2}$ (%)	$\frac{\alpha}{2}$	$1-\frac{\alpha}{2}$	$t_{\frac{\alpha}{2}}$ for df = 24
95	5	2.5	0.025	0.975	± 2.064
96	4	2	0.02	0.98	± 2.172
99	1	0.5	0.005	0.995	± 2.797

Note: used `stats.t.ppf()`

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: Given:

Mean (as claimed) = μ = 270 days

Sample mean = $\bar{x}$ = 260 days

Standard deviation s = 90 days

No of samples n = 18

Degrees of freedom = n-1 = 18-1 = 17

t score : $t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$

t score = $t = \frac{260-270}{\frac{90}{\sqrt{18}}}$ = -0.471

p value corresponding to this t score is

```
stats.t.cdf(x = 260, loc = 270, scale = 90/((18)**(1/2)), df=17)
```

>> 0.32167253567098364

or **32.2%**

Thus, the probability that 18 randomly selected bulbs would have an average life of no more than 260 days is ~ **32.2 %**