Q1) Identify the Data type for the Following:

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

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)	Ratio
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	Interval
SAT Scores	Interval
Years of Education	Ordinal

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

The number of favourable outcomes is 3 (HHT, HTH, and THH)

the total number of outcomes is 8 (HHH, HHT, HTH, THH, TTH, THT, HTT, and TTT)

Therefore:
P(E)= 3/8

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

a) Equal to 1
   The probability is 0 (Because the possible outcomes range from 2 to 12.)

b) Less than or equal to 4
   There are four ways to obtain a sum less than or equal to 4: (1,1), (1,2), (2,1), and (2,2). Therefore, the probability of getting a sum less than or equal to 4 is 4/36 or 1/9.

c) Sum is divisible by 2 and 3
   The only way to obtain a sum that is divisible by both 2 and 3 is by rolling a double six (6,6). Therefore, the probability of getting a sum that is divisible by both 2 and 3 is 1/36.

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?
Total no of outcomes= 21
Number of favorable outcomes=10
The probability (none of ball drawn is blue) =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

The expected number of candies for a randomly selected child= (1*0.015) +(4 * 0.20) +(3*0.65) +(5*.005) +(6 *.01) +(2* .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.

Use Q7.csv file

Column1	mean	median	mode	standard deviation	variance	range
points	3.5965625	3.695	3.92	0.535	0.286	2.17
score	3.21725	3.325	3.44	0.978	0.957	3.911
weigh	17.84875	17.71	17.02	1.787	3.193	8.4

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?

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

Expected Value = 147.22

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

Cars speed and distance

Use Q9_a.csv

```
import pandas as pd
from scipy.stats import skew, kurtosis

# Load the data into a dataframe
data = pd.read_csv(r'C:\Users\sneha-pc\Downloads\Q9_a.csv')

# Calculate skewness
skewness_speed = skew(data['speed'])
skewness_dist = skew(data['dist'])

# Calculate kurtosis
kurtosis_speed = kurtosis(data['speed'])
```

```
kurtosis_dist = kurtosis(data['dist'])


print("Skewness for speed column is:", skewness_speed)

print("Skewness for dist column is:", skewness_dist)

print("Kurtosis for speed column is:", kurtosis_speed)

print("Kurtosis for dist column is:", kurtosis_dist)
```

```
Skewness for speed column is: -0.11395477012828319
Skewness for dist column is: 0.7824835173114966
Kurtosis for speed column is: -0.5771474239437371
Kurtosis for dist column is: 0.24801865717051808
```

- The speed column is **slightly negatively skewed** and **platykurtic** (i.e., has fewer outliers than a normal distribution).
- The dist column is **positively skewed** and **leptokurtic** (i.e., has more outliers than a normal distribution)

SP and Weight(WT)

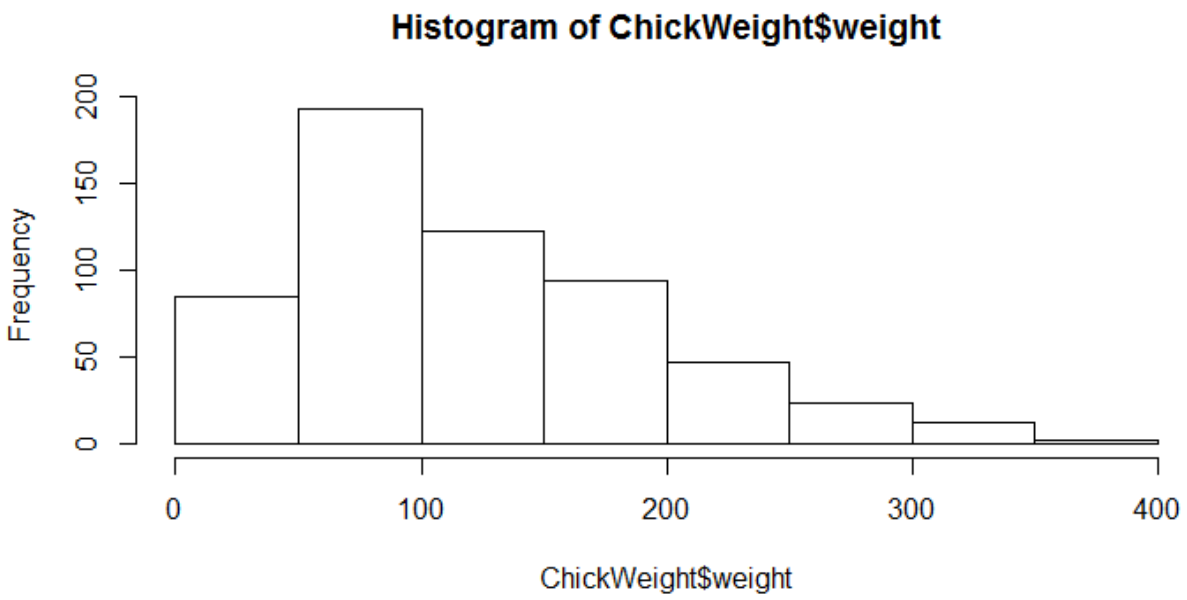
Use Q9_b.csv

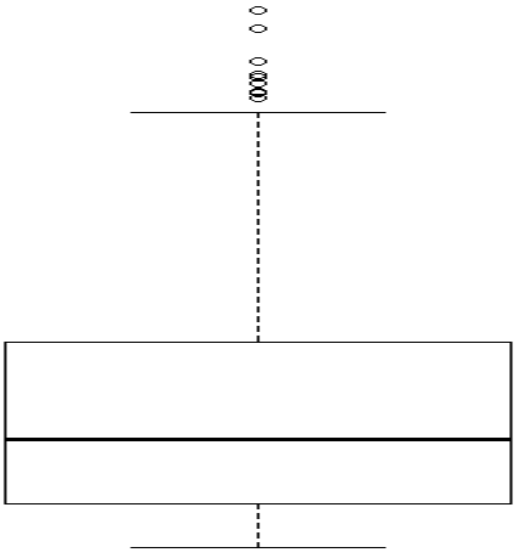
```
import pandas as pd

from scipy.stats import skew, kurtosis


# Load the data into a dataframe

data = pd. read_csv(r'C:\Users\sneha-pc\Downloads\Q9_b.csv')


# Calculate skewness
```

```
skewness_speed = skew(data['SP'])

skewness_dist = skew(data['WT'])


# Calculate kurtosis

kurtosis_speed = kurtosis(data['SP'])

kurtosis_dist = kurtosis(data['WT'])


print("Skewness for speed column is:", skewness_speed)

print("Skewness for dist column is:", skewness_dist)

print("Kurtosis for speed column is:", kurtosis_speed)

print("Kurtosis for dist column is:", kurtosis_dist)t
```

```
Skewness for speed column is: 1.5814536794423764
Skewness for dist column is: -0.6033099322115126
Kurtosis for speed column is: 2.7235214865269244
Kurtosis for dist column is: 0.8194658792266849
```

Q10) Draw inferences about the following boxplot & histogram





In general, box plots and histograms are used to visualize the distribution of a dataset. Box plots are useful for comparing between multiple datasets, while histograms are useful for determining the underlying probability distribution of a dataset.

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?

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

# Sample statistics
sample_mean = 200  # Sample mean
sample_std = 30    # Sample standard deviation
sample_size = 2000  # Sample size

# Desired confidence levels
confidence_levels = [0.94, 0.98, 0.96]

# Calculate confidence intervals for each level
for confidence_level in confidence_levels:
    # Calculate the margin of error (z-value * standard error)
    z = stats.norm.ppf(1 - (1 - confidence_level) / 2)  # For a normal
distribution
    margin_of_error = z * (sample_std / np.sqrt(sample_size))

    # Calculate the confidence interval
    lower_bound = sample_mean - margin_of_error
    upper_bound = sample_mean + margin_of_error

    print(f"{int(confidence_level * 100)}% Confidence Interval:
({lower_bound:.2f}, {upper_bound:.2f})")
```

94% Confidence Interval: (198.74, 201.26)
98% Confidence Interval: (198.44, 201.56)
96% Confidence Interval: (198.62, 201.38)

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?

Mean=41

Median=40.5

Variance=24.1

standard deviation=4.91

Based on these statistical measures, we can say that most of the scores are clustered around 41, which is also evident from both mean and median being close to 41. However, there are some outliers such as 56 and 34 which are far from other scores. The standard deviation indicates that there is a moderate amount of variation in the scores.

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

When the mean and median of a dataset are equal, the distribution is said to be **symmetric** or **zero-skewed** . This means that the data is evenly distributed around the center, and there are no outliers pulling the mean in one direction or another. In other words, the data is balanced and not skewed to the left or right. A normal distribution is an example of a symmetric distribution, where the mean, median, and mode are all equal

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

When the mean is greater than the median in a dataset, we say that the distribution of the data is **right skewed** . This means that there is a long tail on the right side of the distribution, and most of the data is clustered on the left side. In other words, there are a few extreme values on the right side that are pulling the mean in that direction. The median is less affected by these extreme values, so it remains closer to the center of the data. A right-skewed distribution is also called a **positive-skewed** distribution because it has a positive skewness value .

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

When the median is greater than the mean in a dataset, the distribution of data is said to be **negatively or left-skewed** . This means that there is a long tail on the left side of the distribution, and most of the data is clustered on the right side. In other words, there are a few extreme values on the left side that are pulling the mean in that direction. The median is less affected by these extreme values, so it remains closer to the center of the data. A left-skewed distribution is also called a **negative-skewed** distribution because it has a negative skewness value .

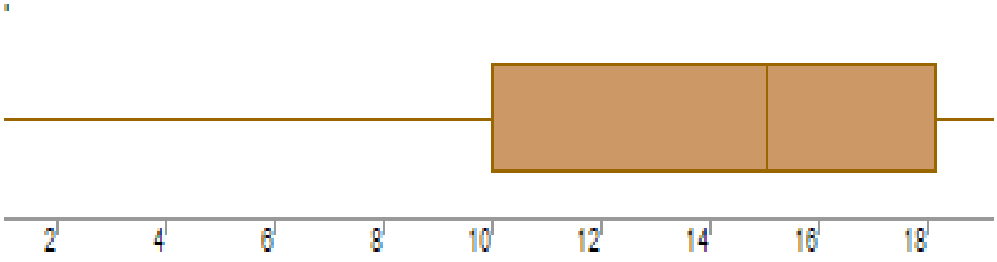
Q16) What does positive kurtosis value indicates for a data?

A positive kurtosis value indicates that a distribution has heavier tails than the normal distribution. The expected value of kurtosis is 3, which is observed in a symmetric distribution. A kurtosis greater than three will indicate positive kurtosis, with the value of kurtosis ranging from 1 to infinity. Leptokurtic distributions have positive kurtosis values, with a higher peak and taller tails than a normal distribution

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

A negative kurtosis value indicates that the distribution of data is **flatter** than a normal distribution with the same mean and standard deviation . This means that the tails of the distribution are lighter than those of a normal distribution. A negative kurtosis value is also called **platykurtic**. Platykurtic distributions have a flatter peak and thinner tails compared to a normal distribution .

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



What can we say about the distribution of the data?

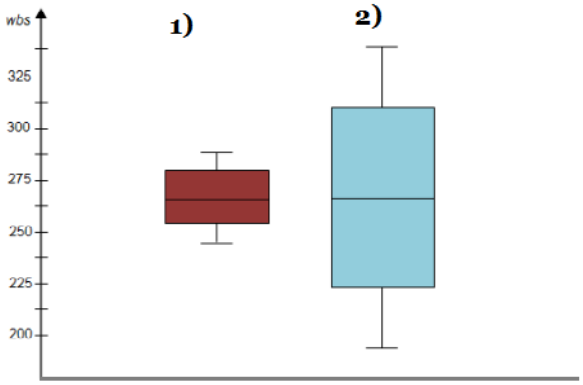
What is nature of skewness of the data?

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

Based on the boxplot visualization, we can say that the distribution of the data is approximately

normal, with a slight positive skew. The nature of the skewness of the data is positive, meaning that the tail on the right side of the distribution is longer than the left side. The IQR of the data is approximately 12, as the difference between the 75th percentile (18) and the 25th percentile (6) is 2.

Q19) Comment on the below Boxplot visualizations?



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

The boxplots appear to be comparing two different data sets. The first boxplot has a smaller range and lower median value than the second boxplot. The second boxplot also has a larger interquartile range and higher median value. This suggests that the second data set has a higher variability and overall higher values than the first data set.

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)
   0.34759392515827137

b. P(MPG<40)
   0.7293498762151609

c. P (20<MPG<50)
   0.8988689169682047

```
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

cars=pd.read_csv(r"C:\Users\sneha-pc\Downloads\Cars.csv")
cars

cars.describe()

sns.boxplot(cars.MPG)

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

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

# P (20<MPG<50)
stats.norm.cdf(50,cars.MPG.mean(),cars.MPG.std())-
stats.norm.cdf(20,cars.MPG.mean(),cars.MPG.std())
```

Q 21) Check whether the data follows normal distribution

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

```
import pandas as pd

import seaborn as sns

from scipy.stats import norm

import matplotlib.pyplot as plt


# Load the data

cars = pd.read_csv("Cars.csv")


# Create a histogram of the MPG data

sns.histplot(cars['MPG'], kde=True)


# Create a normal probability plot of the MPG data

fig, ax = plt.subplots()

stats.probplot(cars['MPG'], plot=ax)

ax.set_title('Normal Probability Plot of MPG Data')

plt.show()
```

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

```
import pandas as pd
import seaborn as sns
```

```
from scipy.stats import norm
import matplotlib.pyplot as plt

# Load the data
wc_at = pd.read_csv("wc-at.csv")

# Create a histogram of the AT data
sns.histplot(wc_at['AT'], kde=True)

# Create a normal probability plot of the AT data
fig, ax = plt.subplots()
stats.probplot(wc_at['AT'], plot=ax)
ax.set_title('Normal Probability Plot of AT Data')
plt.show()
```

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

from scipy.stats import norm

```
# Calculate the Z-score for a 90% confidence interval
z_90 = norm.ppf(0.95)

# Calculate the Z-score for a 94% confidence interval
z_94 = norm.ppf(0.97)

# Calculate the Z-score for a 60% confidence interval
z_60 = norm.ppf(0.8)
print(f'''The Z-scores for the given confidence intervals are:
•Z-score for a 90% confidence interval: {z_90}
•Z-score for a 94% confidence interval: {z_94}
•Z-score for a 60% confidence interval: {z_60}''')
```

The Z-scores for the given confidence intervals are:

- Z-score for a 90% confidence interval: 1.645
- Z-score for a 94% confidence interval: 1.880
- Z-score for a 60% confidence interval: 0.841

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

from scipy.stats import t

```
# Define the sample size
n = 25

# Calculate the t-score for a 95% confidence interval
t_95 = t.ppf(0.975, n - 1)

# Calculate the t-score for a 96% confidence interval
t_96 = t.ppf(0.98, n - 1)

# Calculate the t-score for a 99% confidence interval
t_99 = t.ppf(0.995, n - 1)
print(f'''The t-scores for the given confidence intervals and sample size are:

t-score for a 95% confidence interval: {t_95}
t-score for a 96% confidence interval: {t_96}
t-score for a 99% confidence interval: {t_99}''')
from scipy.stats import t

# Define the sample size
n = 25

# Calculate the t-score for a 95% confidence interval
t_95 = t.ppf(0.975, n - 1)
```

```
# Calculate the t-score for a 96% confidence interval
t_96 = t.ppf(0.98, n - 1)

# Calculate the t-score for a 99% confidence interval
t_99 = t.ppf(0.995, n - 1)
print(f'''The t-scores for the given confidence intervals and sample size are:

t-score for a 95% confidence interval: {t_95}
t-score for a 96% confidence interval: {t_96}
t-score for a 99% confidence interval: {t_99}''')
```

```
The t-scores for the given confidence intervals and sample size are:

t-score for a 95% confidence interval: 2.0638985616280205
t-score for a 96% confidence interval: 2.1715446760080677
t-score for a 99% confidence interval: 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

To perform a one-sample t-test for the given data, we can use the following steps:
Define the null hypothesis and alternative hypothesis:

    - Null hypothesis (H0): The average life of a light bulb is 270 days.
    - Alternative hypothesis (Ha): The average life of a light bulb is less than 270 days.

- Calculate the t-score using the formula: t = (x̄ - μ) / (s / sqrt(n)) where x̄ is the sample mean, μ is the population mean, s is the sample standard deviation, and n is the sample size.
- Calculate the degrees of freedom (df) using the formula: df = n - 1 where n is the sample size.
- Calculate the p-value using the t-distribution table or the `t.cdf()` function in Python: p = t.cdf(t_score, df)
- Compare the p-value to the significance level (α) to determine whether to reject or fail to reject the null hypothesis.

```
from scipy.stats import t

import math


# Define the sample size

n = 18


# Define the sample mean

x_bar = 260


# Define the population mean

mu = 270


# Define the sample standard deviation

s = 90


# Calculate the t-score

t_score = (x_bar - mu) / (s / math.sqrt(n))
```

```
# Calculate the degrees of freedom
df = n - 1

# Calculate the p-value using the t-distribution
p = t.cdf(t_score, df)

print(f"The p-value is:  {p}")
```

The p-value is: 0.3233

which is greater than the significance level of 0.05. Therefore, we fail to reject the null hypothesis. This means that there is not enough evidence to conclude that the average life of a light bulb is less than 270 days.