In [1]:
```python
from IPython import display
display.Image("1.png")
```

Out[1]:

Q1. What are the three measures of central tendency?

Ans. The three measures of central tendency are:

1. Mean: The mean is the sum of all values in a dataset divided by the total number of values. It represents the average value of a dataset.
2. Median: The median is the middle value of a dataset when the values are arranged in ascending order. If there are an even number of values, the median is the average of the two middle values.
3. Mode: The mode is the value that occurs most frequently in a dataset. If there are multiple values with the same highest frequency, the dataset is said to have multiple modes.

In [2]:
```python
from IPython import display
display.Image("2.png")
```

Out[2]:

Q2. What is the difference between the mean, median, and mode? How are they used to measure the central tendency of a dataset?

In [1]:
```python
from IPython import display
display.Image("!.png")
```

Out[1]:

Measure of Central Tendency	Calculation	Sensitivity to Extreme Values	Usefulness	How they are used to measure central tendency
Mean	Sum of all values / Total number of values	Sensitive	When the data is normally distributed or symmetrical	It represents the arithmetic average of the dataset
Median	Middle value when data is ordered from lowest to highest or highest to lowest	Robust	When the data has extreme values or is skewed	It represents the value that splits the dataset in half
Mode	Value that occurs most frequently in a dataset	Insensitive	When we want to know which value occurs most frequently in the data	It represents the most common value in the dataset

In [3]:
```python
from IPython import display
display.Image("3.png")
```

Out[3]:

Q3. Measure the three measures of central tendency for the given height data:

[178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]

In [16]:
```python
import numpy as np
from scipy.stats import mode
```

In [21]:
```python
height = [178, 177, 176, 177, 178.2, 178, 175, 179, 180, 175, 178.9, 176.2, 177, 172.5, 178, 176.5]
mean  = np.mean(height)
print("mean of height: ", mean)
```

```
mean of height:  177.01875
```

In [22]:
```python
median = np.median(height)
print("median of height: ",median)
```

```
median of height:  177.0
```

In [26]:
```python
from scipy.stats import mode
mode = mode(height, keepdims=False)
print("mode of height: ", mode)
```

```
mode of height:  ModeResult(mode=177.0, count=3)
```

In [4]:
```python
from IPython import display
display.Image("4.png")
```

Out[4]:

Q4. Find the standard deviation for the given data:

[178,177,176,177,178.2,178,175,179,180,175,178.9,176.2,177,172.5,178,176.5]

In [27]:
```python
data = [178, 177, 176, 177, 178.2, 178, 175, 179, 180, 175, 178.9, 176.2, 177, 172.5, 178, 176.5]
std = np.std(data)
print("standard deviation of data: ", std)
```

```
standard deviation of data:  1.7885814036548633
```

In [5]:
```python
from IPython import display
display.Image("5.png")
```

Out[5]:

Q5. How are measures of dispersion such as range, variance, and standard deviation used to describe the spread of a dataset? Provide an example.

Ans. Measures of dispersion, such as range, variance, and standard deviation, are used to describe how spread out the data in a dataset is. Here's how each measure is used:

1. Range: The range is the difference between the maximum and minimum values in a dataset. It gives an idea of how much the values in the dataset vary from one another. A wider range indicates a more spread-out dataset, while a smaller range indicates a more tightly clustered dataset. For example, consider the following set of data: 10, 20, 30, 40, 50. The range is 50 - 10 = 40, which means the data spans a range of 40 units.

data = [10,20,30,40,50] range = max(data)-min(data)

= 50 - 10

= 40

1. Variance: Variance is a measure of how much the values in a dataset deviate from the mean. It is calculated by taking the sum of the squared differences between each value and the mean, divided by the total number of values. A higher variance indicates that the data is more spread out, while a lower variance indicates that the data is more tightly clustered around the mean. For example, consider the following set of data: 10, 20, 30, 40, 50. The mean is 30. The variance is calculated as follows:

variance = [(10 - 30)^2 + (20 - 30)^2 + (30 - 30)^2 + (40 - 30)^2 + (50 - 30)^2] / 5

```
= 200 / 5

= 40
```

1. Standard deviation: The standard deviation is the square root of the variance and is expressed in the same units as the data. It is a more intuitive measure of dispersion because it is in the same units as the data. A higher standard deviation indicates that the data is more spread out, while a lower standard deviation indicates that the data is more tightly clustered around the mean. For example, using the same set of data as above, the standard deviation is calculated as follows:

standard deviation = sqrt(variance)

```
= sqrt(40)

= 6.32 (approx.)
```

In [6]:
```python
from IPython import display
display.Image("6.png")
```

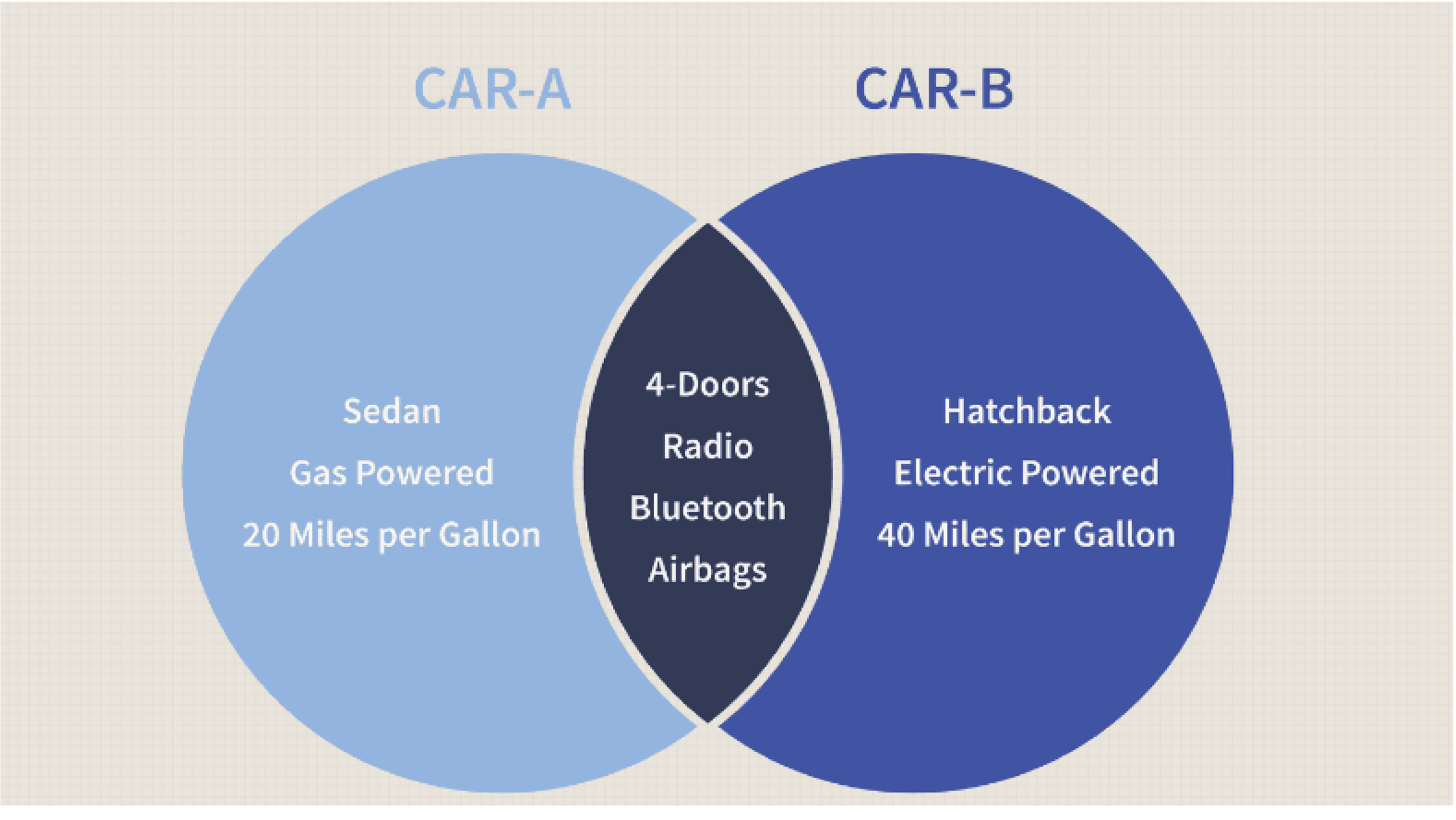
Out[6]:

Q6. What is a Venn diagram?

Ans. A Venn diagram is a graphical representation of sets or groups, which shows all possible logical relations between them. The diagram consists of overlapping circles or other shapes, with each circle representing a set, and the overlapping regions representing the intersection of the sets. The purpose of a Venn diagram is to visually demonstrate how different sets or groups relate to one another and how they overlap or differ. Venn diagrams are commonly used in mathematics, statistics, logic, and computer science to illustrate concepts such as set theory, probability, and logic operations. They can also be used in other fields to illustrate relationships between different categories or concepts.

In [28]:
```python
from IPython import display
display.Image("#.png")
```

Out[28]:



In [8]:
```python
from IPython import display
display.Image("7.png")
```

Out[8]:

Q7. For the two given sets A = (2,3,4,5,6,7) & B = (0,2,6,8,10). Find:

(i) A ∩ B

(ii) A ∪ B

In [29]:
```python
#intersection
A = {2, 3, 4, 5, 6, 7}
B = {0, 2, 6, 8, 10}
print("intersection of A and B: ", A.intersection(B) )
```

```
intersection of two sets:  {2, 6}
```

In [30]:
```python
print("union of A and B: ", A.union(B))
```

```
union of A and B:  {0, 2, 3, 4, 5, 6, 7, 8, 10}
```

In [9]:
```python
from IPython import display
display.Image("8.png")
```

Out[9]:

Q8. What do you understand about skewness in data?

Ans. Skewness of data is a measure of the asymmetry of a probability distribution or dataset. It describes the extent to which a dataset is skewed or "lopsided" relative to a normal distribution. A distribution is said to be skewed if one of its tails is longer or heavier than the other, causing it to deviate from a perfectly symmetrical shape.

1. Positive skewness occurs when the tail on the right side of the distribution is longer or heavier than the tail on the left side. This indicates that the dataset has more high values and fewer low values.
2. Negative skewness occurs when the tail on the left side of the distribution is longer or heavier than the tail on the right side. This indicates that the dataset has more low values and fewer high values.

Skewness can be measured using the skewness statistic. A skewness value of zero indicates that the dataset is perfectly symmetrical, while a positive or negative value indicates that the dataset is skewed in the corresponding direction. Skewness can be an important consideration when interpreting data, as it can affect the accuracy of statistical tests and the validity of certain assumptions about the data.

In [10]:
```python
from IPython import display
display.Image("9.png")
```

Out[10]:

Q9. If a data is right skewed then what will be the position of median with respect to mean?

Ans. If a data is right skewed, then the median will be less than the mean. In a right-skewed distribution, the tail of the distribution is longer on the right side, and this elongation pulls the mean to the right of the median. Therefore, the mean is greater than the median in a right-skewed distribution.

In [11]:
```python
from IPython import display
display.Image("10.png")
```

Out[11]:

Q10. Explain the difference between covariance and correlation. How are these measures used in statistical analysis?

Ans. 1. Covariance : Covariance measures the extent to which two variables vary together. Specifically, it measures the degree to which two variables deviate from their means in similar ways. Mathematically, the covariance between two variables X and Y is calculated as follows:

cov(X,Y) = E[(X - E[X])(Y - E[Y])]

The value of the covariance can range from negative infinity to positive infinity. A positive covariance indicates that the two variables tend to move in the same direction, while a negative covariance indicates that the two variables tend to move in opposite directions. A covariance of zero indicates that the two variables are uncorrelated.

1. Correlation : Correlation, on the other hand, is a standardized measure of the linear relationship between two variables. Correlation takes on values between -1 and 1, where a correlation of 1 indicates a perfect positive relationship, a correlation of -1 indicates a perfect negative relationship, and a correlation of 0 indicates no linear relationship. The correlation coefficient between two variables X and Y is calculated as follows:

corr(X,Y) = cov(X,Y) / (SD[X]*SD[Y])

Correlation is a more commonly used measure than covariance, because it is standardized and easier to interpret. Correlation can also be used to compare the strength of the relationship between different pairs of variables on a common scale.

In [12]:
```python
from IPython import display
display.Image("11.png")
```

Out[12]:

Q11. What is the formula for calculating the sample mean? Provide an example calculation for a dataset.

Ans. The formula for calculating the sample mean is:

$\bar{x} = (x_1 + x_2 + ... + xn) / n$

where x_1, x_2, ..., xn are the values in the dataset, and n is the number of values in the sample.

Consider below 5 Values were sampled from a larger dataset Sample: 4, 6, 2, 9, 5

$\bar{x} = (4 + 6 + 2 + 9 + 5) / 5$

$\bar{x} = 26 / 5$

$\bar{x} = 5.2$

In [13]:
```python
from IPython import display
display.Image("12.png")
```

Out[13]:

Q12. For a normal distribution data what is the relationship between its measure of central tendency?

Ans. For a normal distribution, the three measures of central tendency (mean, median, and mode) are all equal. This is because a normal distribution is symmetric around the mean, so the median (which is the middle value when the data is arranged in order) is also the mean. Additionally, the mode (which is the most frequent value in the distribution) is also equal to the mean and median in a normal distribution.

Therefore, in a normal distribution, the mean, median, and mode are all the same value and represent the center of the distribution. This makes the mean a reliable measure of central tendency for normal data.

In [14]:
```python
from IPython import display
display.Image("13.png")
```

Out[14]:

Q13. How is covariance different from correlation?

Ans. The main difference between covariance and correlation is that covariance is an unstandardized measure that reflects the direction and magnitude of the relationship between two variables, while correlation is a standardized measure that reflects only the strength and direction of the linear relationship between two variables.

In [15]:
```python
from IPython import display
display.Image("14.png")
```

Out[15]:

Q14. How do outliers affect measures of central tendency and dispersion? Provide an example.

Ans. Outliers can have a significant impact on measures of central tendency and dispersion, particularly the mean and standard deviation.

In terms of central tendency, outliers can pull the mean away from the center of the data, making it a less representative measure of the typical value in the dataset. The effect of outliers on the median, however, is less pronounced, as the median is less affected by extreme values.

In terms of dispersion, outliers can also have a significant impact on the standard deviation, which is a measure of the spread of the data around the mean. Because the standard deviation is influenced by the distance between each value and the mean, outliers that are far from the center of the data can cause the standard deviation to be larger than it would be without the outlier.

Example:

In [23]:
```python
import numpy as np

#example without outlier
l = [1,2,3,4,5,1,1]
```

In [24]:
```python
#measure of central tendency
mean = np.mean(l)
median = np.median(l)


#measure of dispersion
std = np.std(l)
```

In [25]:
```python
print("mean: ",mean)
print("median: ", median)


print("standard deviation: ", std)
```

```
mean:  2.4285714285714284
median:  2.0
standard deviation:  1.498298354528788
```

In [26]:
```python
import numpy as np

#example with outlier
v = [1,2,3,4,5,1,1,100]
```

In [27]:
```python
#measure of central tendency
mean1 = np.mean(v)
median1 = np.median(v)

#measure of dispersion
std1 = np.std(v)
```

In [28]:
```python
print("mean: ",mean1)
print("median: ", median1)


print("standard deviation: ", std1)
```

```
mean:  14.625
median:  2.5
standard deviation:  32.29913892041087
```

In [29]:
```python
#observing both example

import pandas as pd
od = {"measures": ["mean", "meadian",  "standard deviation"],
     "without_outlier": [mean, median, std],
     "with_outlier": [mean1, median1, std1]}
```

In [30]:
```python
observing_data = pd.DataFrame(od)
```

In [31]:
```python
observing_data
```

Out[31]:

	measures	without_outlier	with_outlier
0	mean	2.428571	14.625000
1	meadian	2.000000	2.500000
2	standard deviation	1.498298	32.299139

In [32]:
```python
observing_data['differences'] = observing_data['with_outlier'] - observing_data['without_outlier']
```

In [33]:
```python
observing_data
```

Out[33]:

	measures	without_outlier	with_outlier	differences
0	mean	2.428571	14.625000	12.196429
1	meadian	2.000000	2.500000	0.500000
2	standard deviation	1.498298	32.299139	30.800841

In [ ]: