ABHIJEET KUMAR

Q1: Create a 1D array of 9 elements using numpy module and reshape it into 2D array of size 3*3

In [9]:
```
import numpy as np
```

In [21]:
```
import numpy as np

x=np.arange(9)
print(x)
y=x.reshape(3,3)
print(y)
```

```
[0 1 2 3 4 5 6 7 8]
[[0 1 2]
 [3 4 5]
 [6 7 8]]
```

Q2: What will be the output: list3=[x for x in range(10) if x%2==0] print(list3)

In [10]:
```
[0, 2, 4, 6, 8]
```

Out[10]:
```
[0, 2, 4, 6, 8]
```

Q3: Let we have a string text = "PythonProgramming"

Perform Slicing as:

a)Slice the string to obtain first 6 elements.

b)Extract the elements from index 6 to index 13

c)Extract last 5 characters of the string

d)Create a new string by sclicing and concatinating the first 4 and last 3 characters of the string

In [11]:
```
text = "PythonProgramming"
first_six_elements = text[:6]
print("a) First 6 elements:", first_six_elements)
```

```
a) First 6 elements: Python
```

In [16]:
```
t2 = text[6:14]
print("b) From index 6 to index 13:", from_index_6_to_13)
```

```
b) From index 6 to index 13: Programm
```

In [15]:
```
t3 = text[-5:]
print("c) Last 5 characters:", last_five_characters)
```

```
c) Last 5 characters: mming
```

In [17]:
```
new_string = text[:4] + text[-3:]
print("d) New string:", new_string)
```

```
d) New string: Pything
```

Q4: Write a python programme to create two data frames namely d1,d2. Construct d1 utilizing a two dimensional list and create d2 using dictionary

In [22]:
```
import pandas as pd
list1 = [[1,2,3],[4,5,6]]
dic1 = {'Name': ['chery','tej','ram'],
'Age':[20,28,32],
}
d1 = pd.DataFrame(list1)
d2 = pd.DataFrame(dic1)
print(d1)
print(d2)
```

```
   0  1  2
0  1  2  3
1  4  5  6
    Name  Age
0  chery   20
1    tej   28
2    ram   32
```

Q5: Write a python code to discuss in detail about the variance, standard deviation, covariance, correlation

In [23]:
```
import numpy as np
import pandas as pd
data={
'x':[10,15,20,25,30],
'y':[5,10,15,20,25]
}
df=pd.DataFrame(data)
variance_x=np.var(df['x'])
variance_y=np.var(df['y'])
dev_x=np.std(df['x'])
dev_y=np.std(df['y'])
covariance=np.cov(df['x'],df['y'])[0,1]
correlation=np.corrcoef(df['x'],df['y'])[0,1]
print(variance_x)
print(variance_y)
print(dev_x)
print(dev_y)
print(covariance)
print(correlation)
```

```
50.0
50.0
7.0710678118654755
7.0710678118654755
62.5
1.0
```

Q6: Write a python programme to explain the concept of standardization and normalization. Discuss the circumstances under which it is appropiate to utilize these techniques in data processing

In [30]:
```
import numpy as np
from sklearn.preprocessing import StandardScaler, MinMaxScaler
# Sample data
data = np.array([[1, 2, 3],
[4, 5, 6],
[7, 8, 9]])
# Standardization
scaler_std = StandardScaler()
data_std = scaler_std.fit_transform(data)
print("Data after standardization:")
print(data_std)
print()
# Normalization
scaler_norm = MinMaxScaler()
data_norm = scaler_norm.fit_transform(data)
print("Data after normalization:")
print(data_norm)
```

```
Data after standardization:
[[-1.22474487 -1.22474487 -1.22474487]
 [ 0.          0.          0.        ]
 [ 1.22474487  1.22474487  1.22474487]]

Data after normalization:
[[0.  0.  0. ]
 [0.5 0.5 0.5]
 [1.  1.  1. ]]
```

Q8: Explain Following with suitable example

a) Supervised learning and Unsupervised learning

b)Nominal and Ordinal Variable

c)Normalise the following data using min-max normalization by setting min=0, max=1:

1000,2000,3000,9000

a) Supervised Learning: In supervised learning, the algorithm learns from labeled data, meaning each training example consists of input data and its corresponding label or output. The goal is to learn a mapping from inputs to outputs based on the labeled examples provided during training. The algorithm makes predictions on unseen data by generalizing from the labeled examples it has seen Unsupervised Learning: In unsupervised learning, the algorithm learns patterns from unlabeled data. There are no specific output labels associated with the training examples. Instead, the algorithm identifies structure or relationships in the data without explicit guidance. Unsupervised learning algorithms aim to uncover hidden patterns or groupings in the data.

b) Nominal Variable: A nominal variable is a categorical variable with two or more categories that do not have any intrinsic order or ranking

Ordinal Variable: An ordinal variable is a categorical variable with two or more categories that have a natural order or ranking. The categories represent levels of a qualitative attribute that can be ordered from lowest to highest or vice versa. However, the differences between the categories are not necessarily equal.

c) Normalize the following data using min-max normalization by setting min=0, max=1:

Provide a detailed ex

Q9: Provide a detailed explanation of the PCA technique for dimensionality reduction including its methology and application

ans- Principal Component Analysis (PCA) is a powerful technique used in data analysis, particularly for reducing the dimensionality of datasets while preserving crucial information. It does this by transforming the original variables into a set of new, uncorrelated variables called principal components. Here’s a breakdown of PCA’s key aspects: Dimensionality Reduction: PCA helps manage high-dimensional datasets by extracting essential information and discarding less relevant features, simplifying analysis. Data Exploration and Visualization: It plays a significant role in data exploration and visualization, aiding in uncovering hidden patterns and insights. Linear Transformation: PCA performs a linear transformation of data, seeking directions of maximum variance. Feature Selection: Principal components are ranked by the variance they explain, allowing for effective feature selection. Data Compression: PCA can compress data while preserving most of the original information. Clustering and Classification: It finds applications in clustering and classification tasks by reducing noise and highlighting underlying structure. Advantages: PCA offers linearity, computational efficiency, and scalability for large datasets. Limitations: It assumes data normality and linearity and may lead to information loss. Let’s say we have a data set of dimension 300 (n) × 50 (p). n represents the number of observations, and p represents the number of predictors. Since we have a large p = 50, there can be p(p-1)/2 scatter plots, i.e., more than 1000 plots possible to analyze the variable relationship. Wouldn’t it be a tedious job to perform exploratory analysis on this data? In this case, it would be a lucid approach to select a subset of p (p << 50) predictor which captures so much information, followed by plotting the observation in the resultant low-dimensional space. The image below shows the transformation of high-dimensional data (3 dimension) to low-dimensional data (2 dimension) using PCA. Not to forget, each resultant dimension is a linear combination of p features

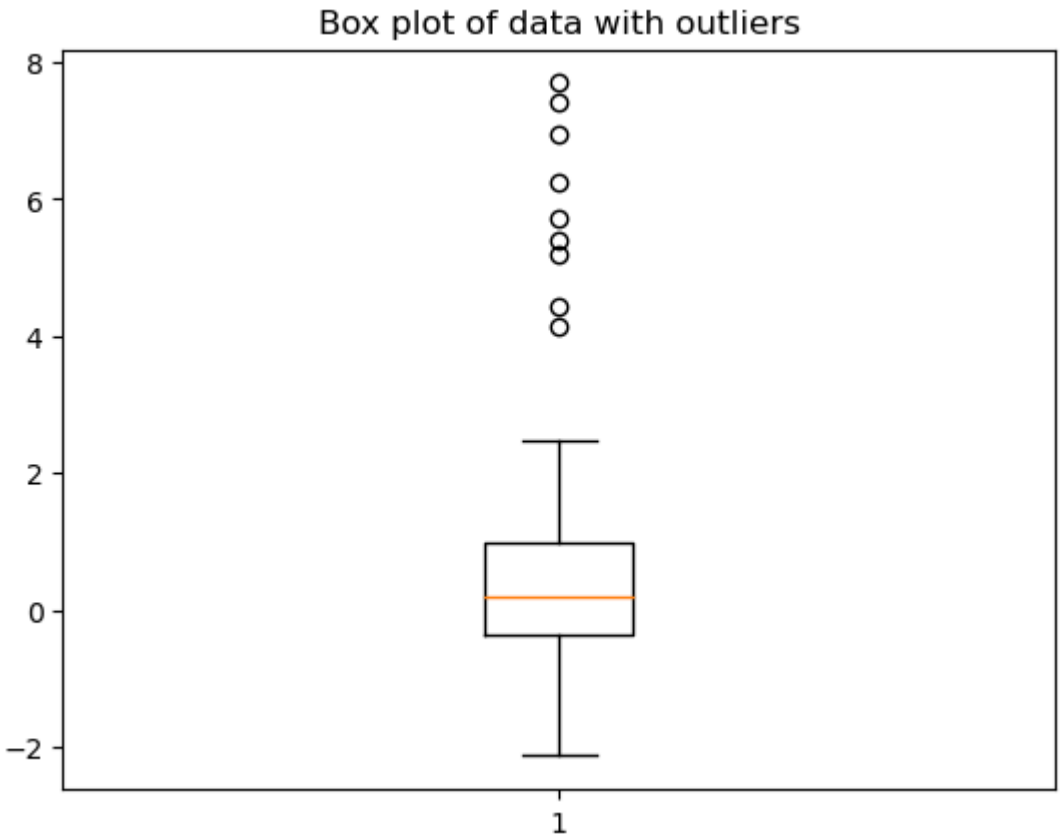
Q10: a)Write a python programme illustrating box plot. Explain how box plot aid in understanding outliers in data

b) Why is the normal distribution important in statistic, and how can it be visualized with a diagram? Define skewness and describe the characteristics of left and right skewed distributions

In [26]:
```
import matplotlib.pyplot as plt
import numpy as np


np.random.seed(10)
data = np.concatenate([np.random.normal(0, 1, 100), np.random.normal(6, 2,


plt.boxplot(data)
plt.title('Box plot of data with outliers')
plt.show()
```



Q-12 Explain concept of hypothesis testing in statistical analysis? Define the p-value in the context of hypothesis testing and explain its significance?

Ans-Hypothesis testing is a form of statistical inference that uses data from a sample to draw conclusions about a population parameter or a population probability distribution. First, a tentative assumption is made about the parameter or distribution. The P value is defined as the probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed. The P stands for probability and measures how likely it is that any observed difference between groups is due to chance. P Value and Statistical Significance: An Uncommon Ground Both the Fisherian and Neyman-Pearson (N-P) schools did not uphold the practice of stating, “P values of less than 0.05 were regarded as statistically significant” or “P-value was 0.02 and therefore there was statistically significant difference.” These statements and many similar statements have criss-crossed medical journals and standard textbooks of statistics and provided an uncommon ground for marrying the two schools. This marriage of inconvenience further deepened the confusion and misunderstanding of the Fisherian and

Neyman-Pearson schools. The combination of Fisherian and N-P thoughts (as exemplified in the above statements) did not shed light on correct interpretation of statistical test of

In [ ]: