School of Advanced Computing & Information Technology

Department of Computer Science, Gujarat University

MCSDE411 (2) Data Analytics (Theory)

Assignment - Unit 1 & 2

1. Explain the differences between categorical and numerical data with examples.
2. Describe the process of handling missing values in a dataset. What are the common techniques used?
3. What is normalization? Explain its importance in data preprocessing and discuss any two normalization techniques.
4. Explain the concept of dimensionality reduction. Why is it important, and what are some common methods used for it?
5. Discuss the differences between univariate, bivariate, and multivariate data with examples.
6. What are outliers? How can they be detected and reduced in a dataset?
7. Explain the differences between cross-sectional and time series data with examples.
8. Discuss the role of Python libraries like Pandas, NumPy, and SciPy in data handling and statistics.
9. What is data preprocessing? Explain its importance and the steps involved in it.
10. Explain the concept of data scales. Discuss the four levels of data measurement (nominal, ordinal, interval, and ratio) with examples.
11. What is difference between graphical and tabular methods.
12. Define Relative frequency.
13. Define Percent frequency distributions.
14. A die is tossed 40 times and lands 6 times on the number 4. What is the relative frequency of observing the die land on the number 4?
15. A coin is tossed 20 times and lands 15 time on heads. What is the relative frequency of observing the coin land on heads?

16. A coin shows heads 15 times having been flipped 40 times. Calculate the relative frequency of the coin showing tails.
    (a) Find the number of times the event occurs.
    (b) Find the number of trials of the experiment.
    (c) Write your answer as a fraction, decimal or percentage.
17. What is stem and leaf display in detail.
18. Explain Ogive.
19. Difference between Matplotlib and Seaborn in data visualization.
20. Write a code of plotting Ogive using matplotlib.
21. Write the difference between Series and Data Frame.
22. What is Quartile explain in brief.
23. Write the difference between Covariance and Correlation**.**
24. A company has collected data on the number of units sold by its five salespeople over the past week. The number of units sold by each salesperson is as follows: 50, 60, 55, 70, and 65. However, after analyzing the data, it was discovered that one of the salespeople incorrectly reported their sales by swapping two digits in the number they submitted. The actual numbers sold by each salesperson should have been: 50, 60, 65, 70, and 55. Calculate the variance and coefficient of variation of the corrected data.
25. Given the following dataset: 10, 12, 15, 18, 21, 24, 100.
    (a) Calculate the mean and standard deviation of the dataset.
    (b) Compute the Z-scores for each data point.
    (c) Identify any outliers in the dataset based on the Z-scores.
26. Write a Python code to load a CSV file using Pandas.
27. What is the output of the following code?

```python
import numpy as np
arr = np.array([1, 2, 3, 4, 5])
print(arr.mean())
```

28. Write a Python code to check for missing values in a Pandas Data Frame.
29. What is the output of the following code?

```python
import pandas as pd
data = {'A': [1, 2, 3], 'B': [4, 5, 6]}
df = pd.DataFrame(data)
print(df.shape)
```

30. Write a Python code to normalize a column in a Pandas Data Frame using Min-Max scaling.
31. What is the output of the following code?

```python
import numpy as np
arr = np.array([[1, 2], [3, 4]])
```

```
print(arr.flatten())
```

32. Write a Python code to drop rows with missing values in a Pandas Data Frame.
33. What is the output of the following code?

```
import pandas as pd
data = {'A': [1, 2, 3], 'B': [4, 5, 6]}
df = pd.DataFrame(data)
print(df['A'].sum())
```

34. Write a Python code to calculate the mean of a NumPy array.
35. What is the output of the following code?

```
import pandas as pd
data = {'A': [1, 2, 3], 'B': [4, 5, 6]}
df = pd.DataFrame(data)
print(df.iloc[1, 1])
```

36. Explain the concepts of mean, median, and mode. How are they calculated, and in what situations is each measure most appropriate?
37. What are percentiles and quartiles? How are they useful in analyzing data distribution? Provide an example to illustrate their application.
38. Define range, interquartile range (IQR), variance, and standard deviation. How do these measures help in understanding data variability?
39. Explain the concept of standard deviation. How does it differ from variance, and why is it considered a more intuitive measure of variability?
40. What is covariance? How is it interpreted, and what are its limitations in measuring the relationship between two variables?
41. Explain the correlation coefficient. How does it differ from covariance, and what does its value indicate about the relationship between two variables?
42. Discuss the interpretation of the correlation coefficient. What do values of -1, 0, and +1 signify in terms of the relationship between variables?
43. What is the empirical rule? How is it used to understand the distribution of data in a normal distribution?
44. Explain the concept of z-scores. How are they calculated, and how do they help in identifying outliers and understanding relative location?
45. What is a box plot? How is it used to detect outliers and analyze the distribution of data?