Data Science Tutorials

Short Questions:

1. Define Data Science and mention any two of its applications.
2. What are the key skills required for a Data Scientist?
3. Briefly explain the concept of Big Data.
4. What is the significance of statistical significance in hypothesis testing?
5. Define the term "Confidence Interval."
6. What are the common sources of data collection?
7. List any two data cleaning techniques and briefly explain one of them.
8. What is the purpose of Exploratory Data Analysis (EDA).
9. What is a histogram, and how is it useful in data visualization?
10. How does NumPy differ from Pandas in Python?
11. What is the role of machine learning in data science?
12. Why is ethical consideration important in data science?
13. What are the ethical considerations in data science?
14. Define data wrangling and its importance.
15. What is the difference between data integration and data transformation?
16. Briefly explain the term "random sampling in data science"
17. What is the purpose of inferential statistics?
18. Define standard error and its significance in data science.
19. What is a p-value in hypothesis testing?
20. Differentiate between Type I and Type II errors.
21. What are the key components of a time series analysis?
22. Define Bayes' theorem with an example.
23. Explain the importance of data preprocessing.
24. What is the difference between structured and unstructured data?
25. How does a scatter plot help in data visualization?
26. What is the difference between bar plots and histograms?
27. Briefly explain the concept of boxplots and their significance.
28. What is the purpose of NumPy in data science?
29. How does Pandas help in handling data?
30. What are the key features of the Seaborn library?

31. Define supervised and unsupervised learning.
32. What is the difference between regression and classification?
33. What is the importance of feature scaling in machine learning?
34. Explain the role of APIs in data collection.
35. What is the significance of data cleaning?
36. Define business analytics and its importance in data science.
37. What is overfitting and underfitting in machine learning, and how can it be prevented?

Long Questions:

1. Explain the importance and applications of data science in modern industries with suitable examples.
2. Discuss the concept of bias and fairness in data science. How can they impact real-world applications?
3. A company's dataset shows a sample mean salary of Rs. 55,000 with a standard deviation of Rs. 5,000 based on 100 employees. Construct a 95% confidence interval for the population mean.
4. Perform a forward and backward fill process in data pre-processing with appropriate example.
5. What are the common methods to detect the missing data, explain any one of them.
6. Explain hypothesis testing and demonstrate it with an example.
7. What is Bayes' Theorem? A company has developed a machine learning model to detect fraud transactions. Based on historical data:
   i. The probability of a transaction being fraud is 2% (i.e., P(F)=0.02
   ii. If a transaction is fraud, the model correctly identifies it 90% of the time (i.e., True Positive Rate, P(T|F) =0.90).
   iii. If a transaction is not fraudulent, the model incorrectly flags it as fraudulent 5% of the time (i.e., False Positive Rate, P(T|NF) =0.05).
   Given that a transaction is flagged as fraud by the model, what is the probability that it is actually fraud?
8. What is regression analysis? Explain with an example of linear regression and its equation.
9. Explain time series analysis and its significance in forecasting.

10. Describe the process of handling missing data in a dataset. Provide Python code for handling missing values.
11. Explain the different techniques of data wrangling and integration.
12. What is Exploratory Data Analysis (EDA)? Write a code that pots a graph depicting number of students who are interested in playing, Basketball, Football, Hockey, Cricket and Table Tennis.
13. A dataset has 200 rows and 5 features. If there are duplicate rows in the data, how would you remove them using Pandas? Write the code to do this.
14. How do you check the business validation for the given data also provide the output achieved after the analysis:

```
data = {'Product': ['A', 'B', 'C'],     'Price': [50, -10, 30]}
```

15. Differentiate between boxplots and histograms. Provide Python code to visualize both.
16. Discuss the significance of machine learning in data science. Provide an example of classification using scikit-learn.
17. Explain different types of probability distributions with examples.
18. Write Python code to read a dataset, clean missing values, and perform a basic analysis.
19. What are the various visualization techniques in Matplotlib and Seaborn? Provide Python examples.
20. Explain NumPy along with its key features.
21. Perform a one sample t- test for the following data set:
    A random sample of 10 products has the weight: [49.5, 48.9,50, 50.5, 49.8,50,48.8.49.9,48.8,50.3]