DATA TOOLKIT

Theory Questions

1. **What is NumPy, and why is it widely used in Python?**
   NumPy is a Python library for numerical computing. It provides support for large multidimensional arrays and matrices, along with a collection of mathematical functions to operate on them. It is widely used for its high performance, support for linear algebra, Fourier transforms, and advanced functions, making it essential for data science and machine learning tasks.

2. **How does broadcasting work in NumPy?**
   Broadcasting is a technique that allows NumPy to perform arithmetic operations on arrays with different shapes by automatically expanding the smaller array to match the dimensions of the larger one.

3. **What is a Pandas DataFrame?**
   A DataFrame is a 2-dimensional, tabular data structure in Pandas, similar to an Excel spreadsheet, with rows and columns. It allows for easy data manipulation, analysis, and visualization.

4. **Explain the use of the `groupby()` method in Pandas.**
   The `groupby()` method is used to group data based on certain criteria, enabling aggregate operations like sum, mean, and count on these groups.

5. **Why is Seaborn preferred for statistical visualizations?**
   Seaborn provides aesthetically pleasing and high-level statistical visualizations. It simplifies complex visualizations and integrates well with Pandas and Matplotlib.

6. **What are the differences between NumPy arrays and Python lists?**
   - NumPy arrays are more efficient in memory and speed than Python lists.
   - They support element-wise operations and broadcasting.
   - Arrays are homogeneous, while Python lists can hold mixed data types.

7. **What is a heatmap, and when should it be used?**
   A heatmap is a graphical representation of data where individual values are represented by varying colors. It is used to visualize matrices, correlations, and density data.

8. **What does the term "vectorized operation" mean in NumPy?**
   Vectorized operations in NumPy are operations applied to arrays without using explicit loops, enabling faster computations by leveraging low-level C code.

9. **How does Matplotlib differ from Plotly?**
   - Matplotlib is a static visualization library ideal for creating basic plots.
   - Plotly is interactive and web-based, offering advanced features like zooming, panning, and 3D plots.

10. **What is the significance of hierarchical indexing in Pandas?**
    Hierarchical indexing allows data to be indexed at multiple levels, making it easier to work with multi-dimensional data within a 2D structure.

11. **What is the role of Seaborn's `pairplot()` function?**
    The `pairplot()` function visualizes relationships between pairs of variables in a dataset, often used for exploratory data analysis.

12. **What is the purpose of the `describe()` function in Pandas?**
    The `describe()` function provides a summary of statistical measures like mean, median, and standard deviation for numerical columns in a DataFrame.

13. **Why is handling missing data important in Pandas?**
    Handling missing data is crucial to ensure data quality and avoid biased or incorrect analysis.

14. **What are the benefits of using Plotly for data visualization?**
    Plotly provides interactive, web-friendly visualizations with advanced capabilities, including 3D plots and dashboards.

15. **How does NumPy handle multidimensional arrays?**
    NumPy supports efficient storage and operations on multidimensional arrays, allowing slicing, reshaping, and broadcasting.

16. **What is the role of Bokeh in data visualization?**
    Bokeh is used for interactive and web-based visualizations, allowing customization and interactivity for large datasets.

17. **Explain the difference between `apply()` and `map()` in Pandas.**
    - `apply()` works on rows/columns of a DataFrame or on Series elements.
    - `map()` is restricted to Series and applies a function element-wise.

18. **What are some advanced features of NumPy?**
    - Linear algebra functions.
    - Random number generation.
    - Broadcasting.
    - Fourier transforms.

19. **How does Pandas simplify time series analysis?**
    Pandas provides built-in time series functionalities like resampling, shifting, and handling datetime data.

20. **What is the role of a pivot table in Pandas?**
    A pivot table summarizes data, enabling aggregate calculations and multi-dimensional analysis.

21. **Why is NumPy's array slicing faster than Python's list slicing?**
    NumPy arrays are stored in contiguous memory blocks, enabling efficient slicing, unlike Python lists.

22. **What are some common use cases for Seaborn?**
    - Correlation heatmaps.
    - Distribution plots.
    - Pairwise relationships using `pairplot`.
    - Visualizing categorical data.