Data Visualization

Quiz 1 – Afternoon

Name: _____ Roll Number: _____

Question 1: Encircle the correct option

1. Seaborn offers various statistical functions for plotting relationships. Which function is ideal for visualizing the linear relationship between two continuous variables?

a) sns.jointplot()
b) sns.regplot()
c) sns.heatmap()
d) sns.barplot()

2. When customizing a scatter plot in Plotly, which attribute controls the marker size of data points?

a) `x` and `y` coordinates
b) `text` attribute
c) `marker` dictionary with a `size` property
d) Legend entries have no effect on marker size.

3. When evaluating the effectiveness of a data visualization, which factor is LEAST important?

a) Accuracy of the data being represented.
b) Clarity of the message being conveyed.
c) Visual appeal and aesthetics of the chart.
d) Ability to generate the visualization quickly without much effort.

4. In Python, which library offers a high level of interactivity for data visualizations beyond static images?

a) Matplotlib
b) Seaborn
c) Pandas
d) Plotly

5. Which of the following statements about pre-processing data for visualization is TRUE?

a) Data cleaning and transformation are unnecessary steps for effective visualization.
b) Outliers should always be removed before creating visualizations.
c) It's crucial to ensure data consistency and handle missing values before plotting.
d) Scaling numerical data is not required for all types of visualizations.

Question 2: Give short answers of the following questions

1. What type of chart is best suited to show trends over time? Line Chart
2. What's the advantage of using a legend in a visualization?
3. In Python, which library is commonly used to read and manipulate CSV files for data visualization? Pandas

Question 3: Scenario Based Question

You are a Data Analyst at a multinational corporation. Your manager has provided you with a dataset named df which contains sales data for the past year. The dataset has the following columns:

Product | Region | Quarter | Sales | Profit

Your task is to perform the following operations using Python libraries pandas, numpy, matplotlib, and seaborn:

- Load the dataset into a pandas DataFrame.
- Use numpy to calculate the total sales and total profit for the year.
- Use pandas to find out which product had the highest sales in each region.
- Use matplotlib to create the following visualizations:
- A line graph showing the trend of sales over the four quarters for each product.
- A bar graph showing the total sales for each product.
- A pie chart showing the proportion of total sales for each region.

Use seaborn to create the following visualizations:

- A boxplot showing the distribution of profit for each product.
- A violin plot showing the distribution of sales for each product.
- A pairplot to visualize the relationship between Sales and Profit for each Product.

Write a Python code for the above scenario. Also add comments in your code to explain your steps.