Question 1

A market research company is analyzing data collected from a survey. The data includes participants' age, gender, income level, and preference for different product categories.

Question: Based on this scenario, what is the primary difference between categorical and numerical data?

A) Categorical data represents qualities, while numerical data represents quantities

B) Categorical data is always ordinal, while numerical data is always continuous

C) Categorical data cannot be analyzed statistically, while numerical data can

D) Categorical data is always represented by numbers, while numerical data is always represented by text

Correct Answer: A

Explanation:

Categorical data represents qualities or categories, such as gender or product preferences, while numerical data represents quantities that can be measured or counted, such as age or income level.

Question 2

A school is tracking the number of students enrolled in each grade and their average test scores across subjects.

Question: What distinguishes continuous data from discrete data in this scenario?

A) Continuous data can take any value within a range, while discrete data has distinct, separate values

B) Continuous data is always categorical, while discrete data is always numerical

C) Continuous data is easier to collect than discrete data

D) Discrete data is more accurate than continuous data

Correct Answer: A

Explanation:

Average test scores are continuous because they can take any value within a range. The number of students enrolled is discrete because it consists of distinct whole numbers.

Question 3

A factory tracks various metrics during production, including the weight of product boxes, the number of defective items produced each day, and the types of materials used in production.

Question: Which metric would likely be measured as continuous data?

A) Number of defective products

B) Types of materials used

C) Weight of product boxes

D) Number of production lines

Correct Answer: C

Explanation:

Weight of product boxes is continuous because it can take any value within a range. The other metrics are either discrete (number-based metrics) or categorical (types of materials).

Question 4

A data visualization team wants to create aesthetically pleasing plots showing trends in customer purchases over time using Python libraries. They are deciding whether to use Seaborn or base Matplotlib (i.e., using only the Matplotlib library) for their visualizations.

Question: Which advantage does Seaborn offer over using base Matplotlib in this context?

- A) Seaborn provides built-in themes for improving the aesthetics of plots

- B) Seaborn allows for statistical visualizations with less code
- C) Seaborn is optimized for handling real-time data visualization
- D) Seaborn replaces Matplotlib and does not rely on it

Correct Answer: A

Explanation:

Seaborn provides **built-in themes and color palettes**, making it easier to create visually appealing plots with **less code** than base Matplotlib. Additionally, Seaborn integrates well with Pandas DataFrames and provides specialized functions for statistical visualizations. However, Matplotlib is still required for full customization, and Seaborn does not replace it entirely.

Question 5

A financial analyst wants to visualize the relationship between stock prices and trading volumes over time using Python's Matplotlib library.

Question: What type of plot should they use to show this relationship?

A) Bar plot

B) Pie chart

C) Scatter plot

D) Box plot

Correct Answer: C

Explanation:

A scatter plot is ideal for visualizing relationships between two numerical variables like stock prices and trading volumes.

Question 6:

An e-commerce company wants to analyze sales trends across multiple product categories over different months using heatmaps.

Question: In what scenario would heatmaps be particularly useful for this analysis?

A) Showing the distribution of sales for a single product category

B) Comparing sales trends between two product categories

C) Visualizing changes in sales across months and categories

D) Displaying exact sales figures for specific products

Correct Answer: C

Explanation:

Heatmaps are ideal for visualizing changes in sales across months and categories as they reveal patterns and trends in multi-dimensional datasets.

Question 7

A health researcher observes that people who exercise regularly tend to have lower blood pressure levels but wants to investigate further.

Question: What does correlation between exercise frequency and blood pressure indicate?

A) One variable causes the other

B) The two variables are statistically related but not necessarily in a causal way

C) The two variables are always independent

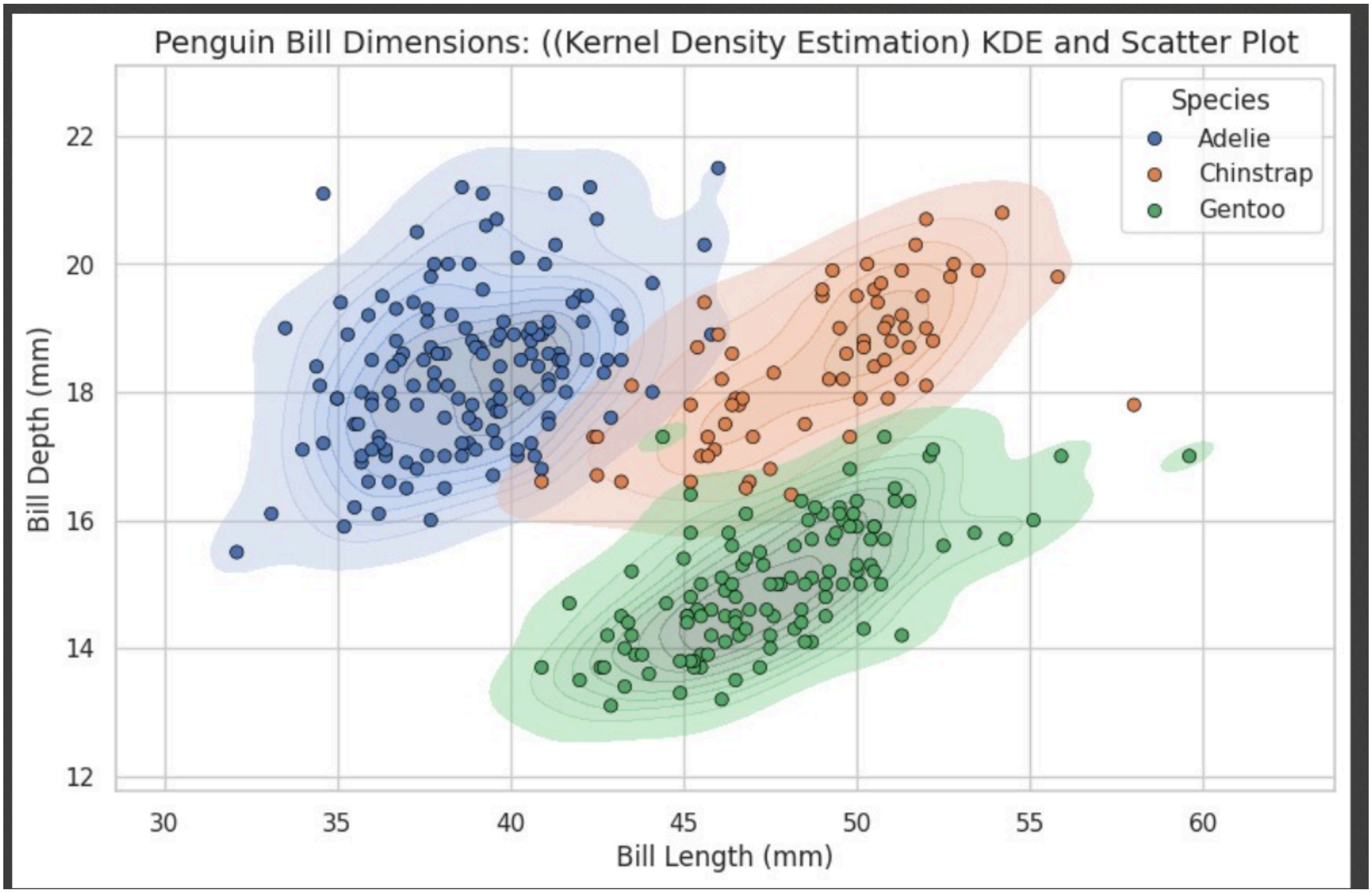
D) One variable can be perfectly predicted from the other

Correct Answer: B

Explanation:

Correlation indicates statistical relationships but does not imply causation. Further investigation is needed to determine if exercise directly affects blood pressure levels.

Question 8



Based on the case study done in class, analyze the attached pic and answer following questions. Question: Which penguin species has lowest mean Bill Length(mm)?

A) Chinstrap

B) Gentoo

C) Adelie

Correct Answer: C

Explanation

The **Adelie** species has the smallest **mean Bill Length (mm)**, as observed from the dataset. In the **KDE plot**, the **Adelie** cluster center's x-axis position is the farthest left, confirming that they have the shortest average bill length compared to Chinstrap and Gentoo penguins.

Question 9:

A sports analytics team wants to compare the running speeds of athletes across different age groups (Under 20, 20–30, 30–40). They want to ensure that individual speed measurements are clearly visible for each age group.

Question: What feature of swarm plots makes them useful in this scenario?

A) They aggregate data into summary statistics like mean and median.

B) They display individual data points without overlap while grouping them by categories.

C) They show trends over time using connected lines.

D) They highlight correlations between numerical variables using color gradients.

Correct Answer: B

Explanation:

Swarm plots display individual data points without overlap while grouping them by categories (e.g., age groups). This makes it easy to see the variation in running speeds within each group.

Question 10:

A real estate company collects data on house prices, square footage, number of bedrooms, and location categories (urban, suburban, rural). The team wants to explore relationships between these variables to understand how they influence house prices.

Question: Why would a pair plot be a good choice for visualizing this dataset?

A) It creates a single plot showing the correlation coefficient between all variables.

B) It generates scatter plots for all numerical variable combinations and includes distributions along the diagonal.

C) It uses bar plots to summarize categorical variables across numerical metrics.

D) It highlights trends over time using line charts for multiple variables.

Correct Answer: B

Explanation:

Pair plots are ideal for visualizing relationships between multiple numerical variables by creating scatter plots for all variable combinations and showing distributions along the diagonal.

Question 11:

Lets say , A businessman is given data of his store sales, he wants to see a graph of how much there has been since last one year. As a data analyst which libraries will you use? **(MSQ)**

A) NumPy
B) Pandas
C) Matplotlib
D) Jupyter Notebook

Correct Answer: A, B, C

Explanation:

- Pandas: For reading the data and understanding the data.
- Numpy: To deal with missing data and statistical calculations.
- Matplotlib : To visualize the data and present it to the business man.
- Jupyter Notebook is not a library.

Question 12:

If a dataset has 1000 data points and 95% of them are redundant, how many unique data points are there? **(NAT)**

Correct Answer: 50

Explanation:

If 95% of the data points are redundant, that means only 5% are unique. So, unique data points = 5% of 1000 = 0.05 * 1000 = 50. This calculation highlights the importance of visualization in managing redundant data.

Question 13:

Your team needs to create a Matplotlib plot comparing weekly coding output and error rates on the same graph with distinct visual styles. Which of the following code snippets correctly customize this dual-line plot? (MSQ)

A)

```
plt.plot(days, code, color='blue', linestyle='-', label='Code Output')
plt.plot(days, errors, color='red', linestyle='--', label='Error Rates')
plt.legend()

plt.show()
```

B)

```
plot(days, code, 'b-', label='Code Output')

plt.plot(days, errors, 'r--', label='Error Rates')
plt.legend()

plt.show()
```

C)

```
plt.plot(days, code, color='blue', style='solid', label='Code Output')

plt.plot(days, errors, color='red', style='dashed', label='Error Rates')
plt.legend()

plt.show()
```

D)

```
plt.plot(days, code, line_color='blue', linestyle='-', label='Code Output')
plt.plot(days, errors, line_color='red', linestyle='--', label='Error Rates')
plt.legend()

plt.show()
```

Correct Answers: A, B

A)

```
plt.plot(days, code, color='blue', linestyle='-', label='Code Output')
plt.plot(days, errors, color='red', linestyle='--', label='Error Rates')
plt.legend()
plt.show()
```

B)

```
plt.plot(days, code, 'b-', label='Code Output')
plt.plot(days, errors, 'r--', label='Error Rates')
plt.legend()
plt.show()
```

Explanation:

This question tests knowledge of Matplotlib’s customization options for plotting two datasets (e.g., coding output and error rates) on the same graph with distinct styles and a legend, a common task in data visualization.

- A) Correct because it uses keyword arguments: color='blue' sets a blue line, linestyle='-' sets a solid line, and label='Code Output' names it for the legend. The second plt.plot() uses color='red'

and linestyle='--' for a dashed red line with label='Error Rates'. plt.legend() adds the legend, and plt.show() displays the plot. This is a standard, explicit approach.
- B) Correct because it uses Matplotlib’s shorthand format: 'b-' means blue solid line (b for blue, - for solid), and 'r--' means red dashed line (r for red, -- for dashed). Labels are provided, and plt.legend() ensures they appear. This is a concise, valid alternative to A.
- C) Incorrect because style='solid' and style='dashed' are not valid parameters; the correct keyword is linestyle (e.g., linestyle='-' or linestyle='--'). This syntax error prevents the plot from rendering as intended.
- D) Incorrect because line_color is not a valid parameter; the correct keyword is color. This would raise an AttributeError in Matplotlib, making the code fail.

Both A and B achieve the goal of a dual-line plot with distinct colors, line styles, and a legend, reflecting practical customization skills needed for visualizing weekly data trends

Question 14:

Given two datasets with values x1= [2,4,6,8,10] and x2=[1,3,5,7,9], what is the mean difference between corresponding points?

Correct Answer: 1

Explanation:

The differences between corresponding points are:

- x1 - x2 = difference
- 2 - 1 = 1
- 4 - 3 = 1
- 6 - 5 = 1
- 8 - 7 = 1
- 10 - 9 = 1 The mean difference is (1 + 1 + 1 + 1 + 1) / 5 = 1. This calculation is a practical example of comparing datasets visually.

Question 15:

Assuming you are working a project and your senior asks you to add grid on the visualized data , what will do to add a gird in Matplotlib plot in Python?

A) Using plt.grid(True)
B) Using plt.add_grid()
C) Using plt.show_grid()
D) Using plt.plot_grid()

Correct Answer: A) Using plt.grid(True)

Explanation:

In Matplotlib, the grid function is used to add grid lines to the plot. The command plt.grid(True) enables the grid, enhancing readability by providing reference lines for data points.

Question 16

Give a scenario, in future you are asked to plot the relation between two types of data by plotting a simple line, as a leaner answer it with respect to this question:
Which of the following code snippets correctly generates a simple line plot with x-values [1, 2, 3] and y-values [4, 5, 6] using Matplotlib?

A)

```
import matplotlib.pyplot as plt
plt.plot([4, 5, 6], [1, 2, 3])
plt.show()
```

B)

```
import matplotlib.pyplot as plt
plt.plot([1, 2, 3], [4, 5, 6])
plt.show()
```

C)

```
import matplotlib as plt
plt.line([1, 2, 3], [4, 5, 6])
plt.show()
```

D)

```
import matplotlib.pyplot as plt
plt.plot([1, 2, 3], [4, 5, 6], [2, 3, 4])
plt.show()
```

Correct Answer: B)

```
import matplotlib.pyplot as plt
plt.plot([1, 2, 3], [4, 5, 6])
plt.show()
```

Explanation:

In Matplotlib, the plt.plot() function takes two arguments: the x-values and y-values, in that order (i.e., plt.plot(x, y)). Option B correctly uses [1, 2, 3] as x-values and [4, 5, 6] as y-values, followed by plt.show() to display the plot.

- A) Incorrect because it reverses the x and y values ([4, 5, 6] as x, [1, 2, 3] as y), which plots the inverse relationship.
- C) Incorrect because import matplotlib as plt is wrong (should be matplotlib.pyplot), and plt.line() is not a valid function (plt.plot() is correct).
- D) Incorrect because plt.plot() does not accept three data lists; it expects only two (x and y), and adding a third causes a syntax error.

Question 17

Some times to create a better understanding of data a data analyst needs to play with colors, cause if all the lines are of the same color how will one know the difference, considering the scenario answer the following question.
What is the correct Matplotlib syntax to change the line colour to red and set the marker style to circles in a line plot?

A)

```
plt.plot(x, y, color='red', marker='o')
```

B)

```
plt.plot(x, y, linecolor='red', markerstyle='circle')
```

C)

```
plt.plot(x, y, 'red', 'o')
```

D)

```
plt.plot(x, y, color='r', marker='circle')
```

Correct Answer: A)

```
plt.plot(x, y, color='red', marker='o')
```

Explanation:

Matplotlib’s plt.plot() allows customization via keyword arguments. The color parameter sets the line color, and marker specifies the marker style. 'red' is a valid color name, and 'o' represents circular markers.

- B) Incorrect because linecolor and markerstyle are not valid parameters; the correct ones are color and marker.
- C) Incorrect because passing 'red' and 'o' as positional arguments without keywords is not the standard syntax (though a shorthand like 'ro' could work, it’s not an option here).
- D) Incorrect because marker='circle' is invalid; 'o' is the correct marker code for circles, and color='r' is a shorthand, but A uses the more explicit color='red', which is preferred for clarity.

Question 18:

A dataset tracks daily coding output over 7 days with values [50, 120, 200, 180, 300, 50, 20]. What is the average number of lines of code written per day?(rounded to 2 decimal places) **(NAT)**

Approach:

1. To find the average, sum all the daily coding output values in the dataset.
2. Divide the total by the number of days and round the result to two decimal places.

Correct Answer: 131.43 (rounded to 2 decimal places)

Explanation:

To find the average lines of code written per day:

- Dataset: [50, 120, 200, 180, 300, 50, 20]
- Sum = 50 + 120 + 200 + 180 + 300 + 50 + 20 = 920
- Number of days = 7
- Average = 920 / 7 ≈ 131.42857 ≈ 131.43 (rounded)
  This calculation demonstrates a basic data analysis task often visualized using plots to identify trends, such as daily coding productivity.