GTSC2143 Machine Learning for Business

Tutorial 3

Please write down your answers in this document and submit it at iSpace by the end of this tutorial.

Activity 1. Data Preparation

1. Load the Dataset
   a) Load the Superstore Sales Dataset using pandas:

```
import pandas as pd
df=pd.read_csv("https://raw.githubusercontent.com/WuCandice/Superstore-Sales-Analysis/refs/heads/main/dataset/Superstore%20Dataset.csv")
```

   b) Display basic information:
      - Dataset shape
      - First 5 rows
      - Column names and data types
   c) Check for any missing values in the dataset

Activity 2. Basic Matplotlib Visualizations

1. Histogram of Sales
   a) Create a histogram of the 'Sales' column with:
      - 20 bins
      - Color: “steelblue”
      - Transparency (alpha): 0.7
      - Proper title, x-label, and y-label
   b) Analysis: Write 2-3 sentences interpreting the distribution of sales values.

2. Box Plot of Profit by Category
   a) Create a box plot comparing “Profit” across different “Category” values using:
      - Color: “lightgreen”
      - Proper title and axis labels
   b) Analysis: Write 2-3 sentences describing the profit distribution differences between categories.

3. Scatter Plot of Sales vs Profit
   a) Create a scatter plot with:
      - X-axis: Sales
      - Y-axis: Profit
      - Color: “coral”
      - Alpha: 0.6
      - Proper title and axis labels
   b) Analysis: Write 2-3 sentences describing the relationship between sales and profit.

4. Bar Chart of Total Sales by Region
   a) Group the data by “Region” and calculate total sales for each region

b) Create a bar chart with:
   - Color: “darkblue”
   - Proper title and axis labels
   - Display values on top of each bar
c) Analysis: Write 2-3 sentences comparing sales performance across regions.

Activity 3. Advanced Matplotlib Visualizations

1. Multi-Panel Figure (2x2 Subplots)
   a) Create a figure with 2x2 subplots containing all four visualizations from Activity 2
   b) Apply consistent styling:
      - Figure size: (15, 12)
      - Tight layout
      - Consistent color scheme
   c) Add a main title for the entire figure

2. Customization Practice
   a) Modify the scatter plot to include:
      - Different colors for each category (use a color map)
      - Legend showing categories
      - Grid lines
   b) Enhance the bar chart with:
      - Different colors for each region
      - Rotated x-axis labels if needed

Activity 4. Seaborn Visualizations

1. Advanced Statistical Plots
   a) Create a pair plot using seaborn for numerical columns:
      - Variables: “Sales”, “Profit”, “Quantity”, “Discount”
      - Use “Category” as hue
      - Apply a consistent color palette
   b) Analysis: Write 2-3 sentences about patterns you observe in the relationships.

2. Joint Plot
   a) Create a joint plot of Sales vs Profit with:
      - Kind: “scatter”
      - Add regression line
      - Use appropriate color
   b) Analysis: Write 2-3 sentences about the correlation and distribution patterns.

3. Heatmap of Correlations
   a) Calculate correlation matrix for numerical columns

b) Create a heatmap using seaborn with:
   - Annotations showing correlation values
   - Appropriate color map (“coolwarm”)
   - Proper title
c) Analysis: Write 2-3 sentences identifying the strongest correlations.

- End of Tutorial 3 -