

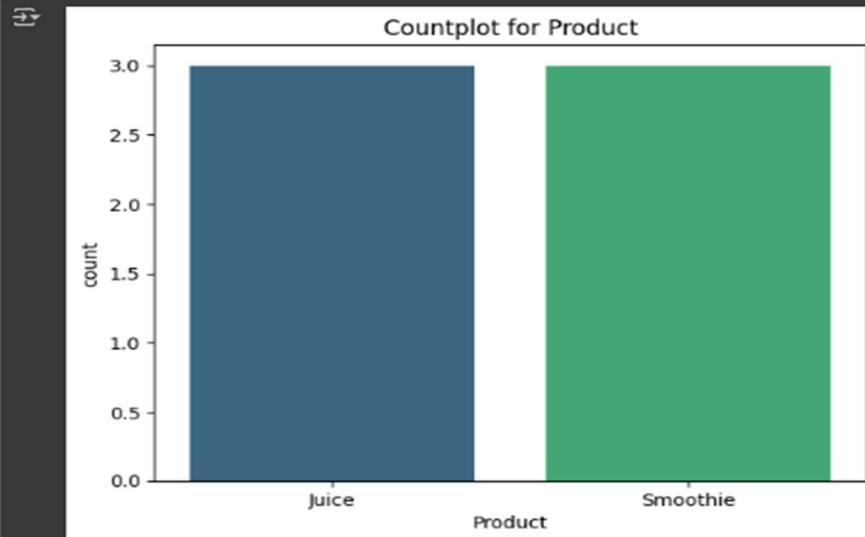
Activity No. 9 Exploring Data Visually	
Course Code: CPE031	Program: BSCPE
Course Title: Visualization and Data Analysis	Date Performed: 10/16/25
Section: CPE21S2	Date Submitted: 10/16/2025
Name: Aicon H. Keliste	Instructor: Maam Sayo
1. Discussion	
<p>Exploratory Data Analysis (EDA) is the process of analyzing datasets to summarize their main characteristics and gain insights before formal modeling. Visual exploration helps identify trends, relationships, and data quality issues.</p> <p>Key Components:</p> <p>Univariate Analysis: Analyzing one variable at a time (e.g., histograms, bar charts).</p> <p>Bivariate Analysis: Exploring relationships between two variables (e.g., scatterplots, crosstabs).</p> <p>Missing Data Analysis: Identifying and handling missing or null values.</p> <p>Time-Series Visualization: Displaying data across time to reveal trends or seasonality.</p> <p>Geospatial Visualization: Mapping data to geographical regions to identify spatial patterns.</p> <p>EDA is both an art and science combining statistical summaries with visual intuition.</p>	
2. Materials and Equipment	
Personal Computer Google Colab	
3. Procedure	
Color and Perception Section 1: Organizing and Exploring Data Section 2: Relationships Between Variables Section 3: Analysis of Missing Data Section 4: Visualizing Time-Series Data Section 5: Visualizing Geospatial Data	
4. Output	

Section 1:

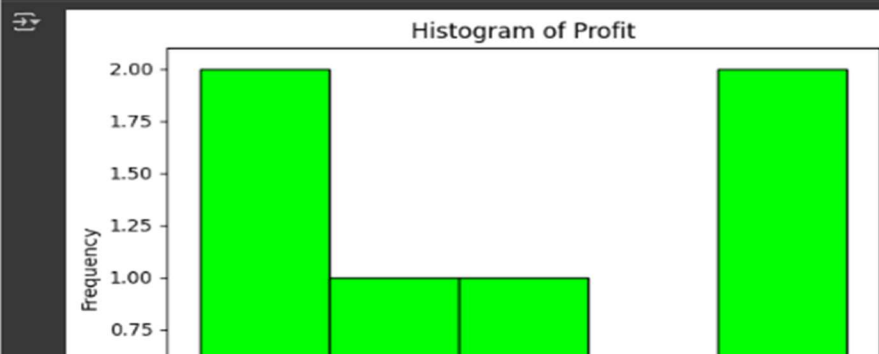
Task:

Create one countplot for Product and one histogram for Profit.
Compare their patterns and interpret what they mean.

```
7] 0s  
#Code here  
sns.countplot(x="Product", data=df, hue="Product", palette="viridis", legend=False)  
plt.title("Countplot for Product")  
plt.show()
```



```
12] 0s  
# Histogram for Profit  
plt.hist(df["Profit"], bins=5, color="lime", edgecolor="black")  
plt.title("Histogram of Profit")  
plt.xlabel("Profit")  
plt.ylabel("Frequency")  
plt.show()
```



Section 2:

Section 2: Relationships Between Variables

```
[ ] # Scatterplot: Sales vs Profit
sns.scatterplot(x="Sales", y="Profit", hue="Region", data=df, palette="coolwarm", s=100)
plt.title("Relationship between Sales and Profit")
plt.show()
```



```
[ ] # Crosstabulation example
pd.crosstab(df["Region"], df["Product"], values=df["Sales"], aggfunc="mean").fillna(0)
```

[↕]

Product			Juice	Smoothie
Region				
East				
			90.0	170.0
North				
			125.0	0.0
South				
			0.0	150.0
West				
			0.0	110.0

Task:

Interpret whether higher sales also mean higher profit.

Which region seems to perform best?

*Answer here

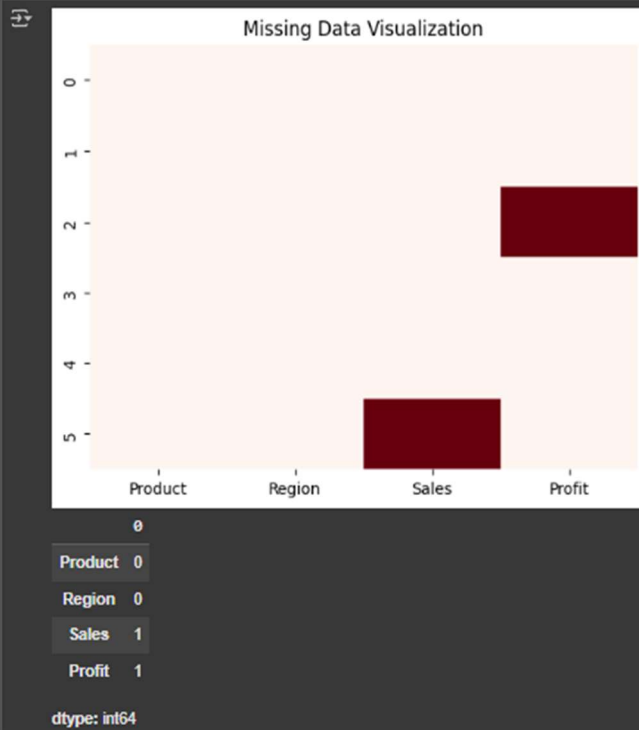
Based on the scatter plot and the table above you can see which regions generate the most Sales and how those Sales relate to Profit

Section 3:

```
[ ] # Introduce missing data
df_missing = df.copy()
df_missing.loc[2, "Profit"] = np.nan
df_missing.loc[5, "Sales"] = np.nan

# Visualize missing data
sns.heatmap(df_missing.isnull(), cbar=False, cmap="Reds")
plt.title("Missing Data Visualization")
plt.show()

# Display missing summary
df_missing.isnull().sum()
```



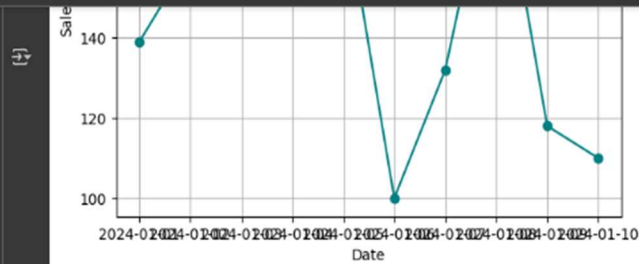
Task:

Describe what you observe in the missing data visualization.
Which variables need attention before analysis?

*Answer here

I observed in the missing data visualization that there are isolated missing values limited to Sales and Profit which means those two numerical variables need attention before analysis.

Section 4:



Task:

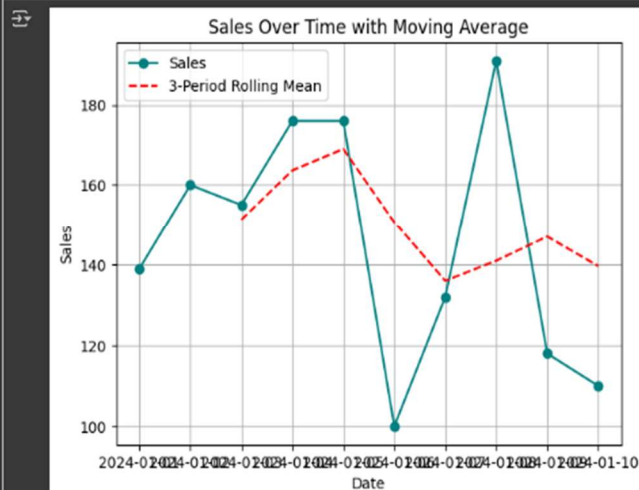
Add a moving average line (rolling mean) to smooth fluctuations.

Hint: Use `ts_df["Sales"].rolling(window=3).mean()`.

[16]

```
# Add a moving average line
ts_df["Sales_RollingMean"] = ts_df["Sales"].rolling(window=3).mean()

# Plot time-series with moving average
plt.plot(ts_df["Date"], ts_df["Sales"], marker="o", color="teal", label="Sales")
plt.plot(ts_df["Date"], ts_df["Sales_RollingMean"], color="red", linestyle="--", label="3-Period Rolling Mean")
plt.title("Sales Over Time with Moving Average")
plt.xlabel("Date")
plt.ylabel("Sales")
plt.grid(True)
plt.legend()
plt.show()
```



Section 5:

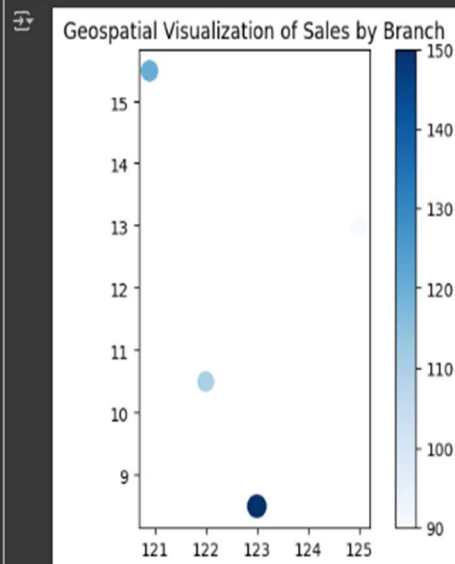
Section 5: Visualizing Geospatial Data

```
[17]
✓ Os
import geopandas as gpd
from shapely.geometry import Point

# Sample coordinates (latitude, longitude)
coords = {
    "Branch": ["North", "South", "East", "West"],
    "Latitude": [15.5, 8.5, 13.0, 10.5],
    "Longitude": [120.9, 123.0, 125.0, 122.0],
    "Sales": [120, 150, 90, 110]
}

geo_df = pd.DataFrame(coords)
geo_df["geometry"] = [Point(xy) for xy in zip(geo_df.Longitude, geo_df.Latitude)]
gdf = gpd.GeoDataFrame(geo_df, geometry="geometry")

# Plot simple map (Philippines outline optional if available)
gdf.plot(column="Sales", cmap="Blues", legend=True, markersize=geo_df["Sales"])
plt.title("Geospatial Visualization of Sales by Branch")
plt.show()
```



Task:

Interpret which branch has the highest sales geographically.

How can such visualization help in business decision-making?

*Answer here

Based on the geospatial visualization I can identify which branch has the highest sales by looking at the size and color of the marker.

Mapping sales by branch reveals spatial patterns of demand that support targeted actions.

Supplementary:

5. Supplementary Activity

Create your own mini exploratory analysis:

1. Choose a dataset (e.g., from Kaggle, or a CSV file you have).
2. Perform:
 - Univariate analysis (1 categorical + 1 quantitative)
 - Bivariate analysis (scatterplot or crosstab)
 - Missing data visualization
 - Time-series or geospatial visualization (choose one)
3. Summarize your key insights using visual interpretation.

```
[3] 20s
from google.colab import drive
drive.mount('/content/drive')
```

Mounted at /content/drive

```
[4] 0s
df = pd.read_csv("/content/drive/MyDrive/VDA/GroceryStoreDataset.csv")
df
```

	MILK, BREAD, BISCUIT
0	BREAD, MILK, BISCUIT, CORNFLAKES
1	BREAD, TEA, BOURNVITA
2	JAM, MAGGI, BREAD, MILK
3	MAGGI, TEA, BISCUIT
4	BREAD, TEA, BOURNVITA
5	MAGGI, TEA, CORNFLAKES
6	MAGGI, BREAD, TEA, BISCUIT
7	JAM, MAGGI, BREAD, TEA
8	BREAD, MILK
9	COFFEE, COCK, BISCUIT, CORNFLAKES
10	COFFEE, COCK, BISCUIT, CORNFLAKES
11	COFFEE, SUGER, BOURNVITA
12	BREAD, COFFEE, COCK
13	BREAD, SUGER, BISCUIT
14	COFFEE, SUGER, CORNFLAKES
15	BREAD, SUGER, BOURNVITA
16	BREAD, COFFEE, SUGER
17	BREAD, COFFEE, SUGER
18	TEA MILK COFFEE CORNFLAKES

Variables Terminal

```

# Prepare data
df_split = df['MILK,BREAD,BISCUIT'].str.split(',', expand=True)

all_items = df_split.stack()
all_items = all_items.reset_index(drop=True)

df_transformed = pd.get_dummies(df_split.stack()).groupby(level=0).sum()
df_transformed = df_transformed.astype(bool)

df['Number_of_Items'] = df_split.count(axis=1)

display(df_transformed.head())
display(df.head())

```

	BISCUIT	BOURNVITA	BREAD	COCK	COFFEE	CORNFLAKES	JAM	MAGGI	MILK	SUGER	TEA
0	True	False	True	False	False	True	False	False	True	False	False
1	False	True	True	False	False	False	False	False	False	False	True
2	False	False	True	False	False	False	True	True	True	False	False
3	True	False	False	False	False	False	False	True	False	False	True
4	False	True	True	False	False	False	False	False	False	False	True

	MILK,BREAD,BISCUIT	Number_of_Items
0	BREAD,MILK,BISCUIT,CORNFLAKES	4
1	BREAD,TEA,BOURNVITA	3
2	JAM,MAGGI,BREAD,MILK	4
3	MAGGI,TEA,BISCUIT	3
4	BREAD,TEA,BOURNVITA	3


```

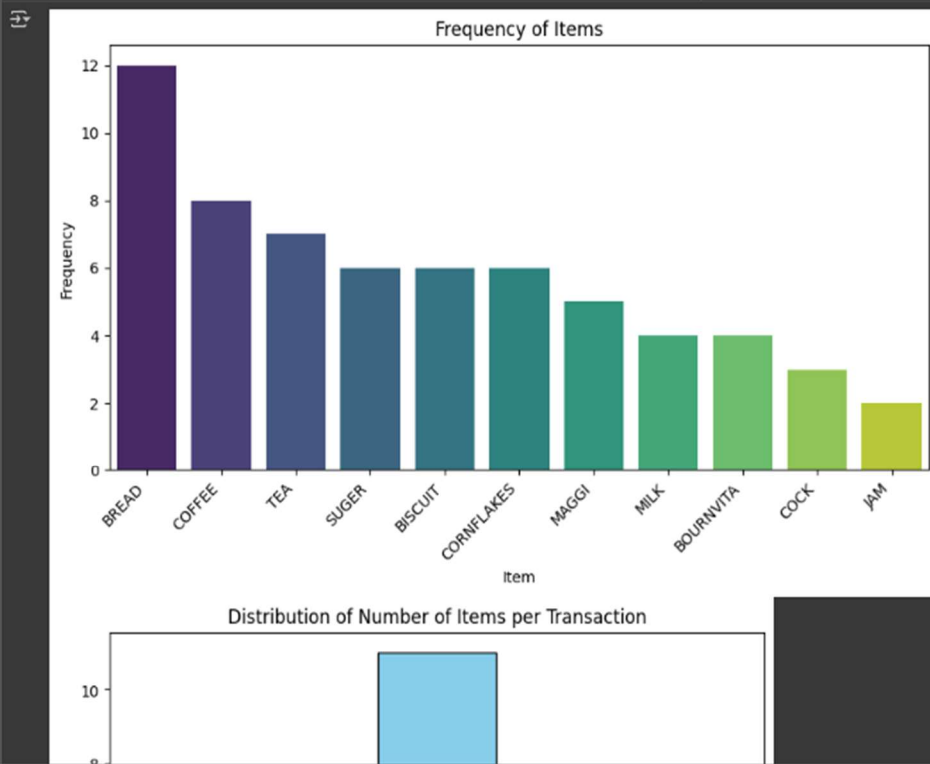
# Univariate analysis

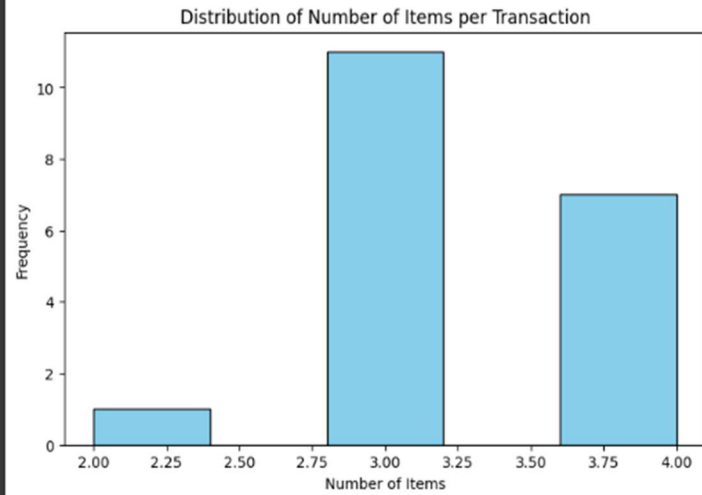
# Categorical variable: Countplot of all items
item_counts = all_items.value_counts()

plt.figure(figsize=(10, 5))
sns.barplot(x=item_counts.index, y=item_counts.values, hue=item_counts.index, palette="viridis", legend=False)
plt.title("Frequency of Items")
plt.xlabel("Item")
plt.ylabel("Frequency")
plt.xticks(rotation=45, ha="right")
plt.show()

# Quantitative variable: Histogram of number of items per transaction
plt.figure(figsize=(8, 5))
plt.hist(df["Number_of_Items"], bins=5, color="skyblue", edgecolor="black")
plt.title("Distribution of Number of Items per Transaction")
plt.xlabel("Number of Items")
plt.ylabel("Frequency")
plt.show()

```

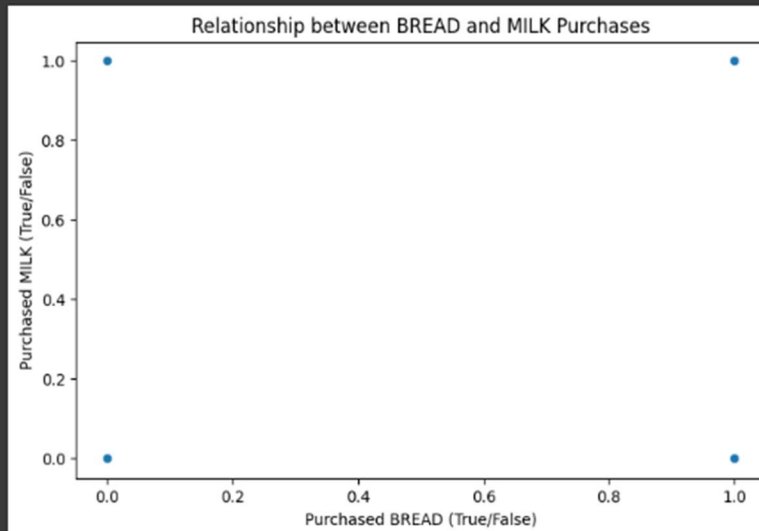




```
# Bivariate analysis (scatterplot or crosstab)

plt.figure(figsize=(8, 5))
sns.scatterplot(x=df_transformed['BREAD'], y=df_transformed['MILK'])
plt.title("Relationship between BREAD and MILK Purchases")
plt.xlabel("Purchased BREAD (True/False)")
plt.ylabel("Purchased MILK (True/False)")
plt.show()

# Crosstabulation: Relationship between two categorical variables (e.g., BREAD and MILK purchase)
crosstab_result = pd.crosstab(df_transformed['BREAD'], df_transformed['MILK'])
print("\nCrosstabulation of BREAD and MILK Purchases:")
print(crosstab_result)
```



Crosstabulation of BREAD and MILK Purchases:

MILK	False	True
BREAD False	6	1
BREAD True	9	3









5. Conclusion/Analysis/Learnings

6. Conclusion/Learnings/Analysis:

In this hands-on activity, I gained more understanding of how visual exploration and analysis techniques can be used for interpreting data analysis. I learned how to use different types of plots to analyze the characteristics of individual variables and the relationships between variables. Lastly, this activity demonstrated the power of data visualization in gaining insights and informing further analysis or decision-making.

6. Assessment Rubric

Lab Activity Rubric							 
Criteria		Ratings					Pts
 SO 7 PI 1 Student Outcome 7.1 Acquire and apply new knowledge from outside sources. threshold: 4.8 pts	6 pts Excellent Educational interests and pursuits exist and flourish outside classroom requirements,knowledge and/or experiences are pursued independently and applies knowledge learned into practice	5 pts Good Educational interests and pursuits exist and flourish outside classroom requirements,knowledge and/or experiences are pursued independently	4 pts Satisfactory Look beyond classroom requirements, showing interest in pursuing knowledge independently	3 pts Unsatisfactory Begins to look beyond classroom requirements, showing interest in pursuing knowledge independently	2 pts Poor Relies on classroom instruction only	1 pts Very Poor No initiative or interest in acquiring new knowledge	6 pts
 SO 7 PI 2 Student Outcome 7.2 Learn independently threshold: 4.8 pts	6 pts Excellent Completes an assigned task independently and practices continuous improvement	5 pts Good Completes an assigned task without supervision or guidance	4 pts Satisfactory Requires minimal guidance to complete an assigned task	3 pts Unsatisfactory Requires detailed or step-by-step instructions to complete a task	2 pts Poor Shows little interest to complete a task independently	1 pts Very Poor No interest to complete a task independently	6 pts
 SO 7 PI 3 Student Outcome 7.3 Critical thinking in the broadest context of technological change threshold: 4.8 pts	6 pts Excellent Synthesizes and integrates information from a variety of sources; formulates a clear and precise perspective; draws appropriate conclusions	5 pts Good Evaluate information from a variety of sources; formulates a clear and precise perspective.	4 pts Satisfactory Analyze information from a variety of sources; formulates a clear and precise perspective.	3 pts Unsatisfactory Apply the gathered information to formulate the problem	2 pts Poor Gather and summarized the information from a variety of sources but failed to formulate the problem	1 pts Very Poor Gather information from a variety of sources	6 pts
 SO 7 PI 4 Student Outcome 7.4 Creativity and adaptability to new and emerging technologies threshold: 4.8 pts	6 pts Excellent Ideas are combined in original and creative ways in line with the new and emerging technology trends to solve a problem or address an issue.	5 pts Good Ideas are creative and adapt the new knowledge to solve a problem or address an issue	4 pts Satisfactory Ideas are creative in solving a problem, or address an issue	3 pts Unsatisfactory Shows some creative ways to solve the problem	2 pts Poor Shows initiative and attempt to develop creative ideas to solve the problem	1 pts Very Poor Ideas are copied or restated from the sources consulted	6 pts
Total Points: 24							