***SHAMPOO SALES DATASET***

*# TASK 1 : FINDING MEAN OF GROUPED DATA*class\_intervals = [(0, 10), (10, 20), (20, 30), (30, 40)]  
frequencies = [5, 8, 15, 7]  
midpoints = [(lower + upper) / 2 for lower, upper in class\_intervals]  
fx = [f \* x for f, x in zip(frequencies, midpoints)]  
mean = sum(fx) / sum(frequencies)  
print(f"Mean of grouped data: {mean}")  
  
  
  
  
*# Task 2: APPLYING DIFFERENT FUNCTIONS ON DATA SET*import pandas as pd  
import numpy as np  
from scipy import stats  
  
*# Load data from CSV*df = pd.read\_csv("shampoo\_sales.csv")  
*# Descriptive statistics*mean = df["Sales"].mean()  
median = df["Sales"].median()  
mode = df["Sales"].mode().tolist()  
variance = df["Sales"].var()  
std\_dev = df["Sales"].std()  
cv = (std\_dev / mean) \* 100  
skewness = stats.skew(df["Sales"])  
kurtosis = stats.kurtosis(df["Sales"])  
z\_scores = stats.zscore(df["Sales"])  
percentiles = {  
 "25th Percentile": np.percentile(df["Sales"], 25),  
 "50th Percentile (Median)": np.percentile(df["Sales"], 50),  
 "75th Percentile": np.percentile(df["Sales"], 75),  
}  
correlation = df["Sales"].corr(pd.Series(range(len(df))))  
  
*# Print results*print("----- Descriptive Statistics -----")  
print(f"Mean: {mean:.2f}")  
print(f"Median: {median}")  
print(f"Mode: {mode}")  
print(f"Variance: {variance:.2f}")  
print(f"Standard Deviation: {std\_dev:.2f}")  
print(f"Coefficient of Variation: {cv:.2f}%")  
print(f"Skewness: {skewness:.2f}")  
print(f"Kurtosis: {kurtosis:.2f}")  
print(f"Correlation Coefficient (with Time): {correlation:.2f}")  
  
print("\n----- Percentiles & Quartiles -----")  
for k, v in percentiles.items():  
 print(f"{k}: {v}")  
  
print("\nZ-scores:")  
print(z\_scores)  
  
  
  
  
  
import matplotlib.pyplot as plt  
import seaborn as sns  
  
*# HISTOGRAM*plt.figure(figsize=(8, 5))  
sns.histplot(df["Sales"], kde=True, color='skyblue', bins=10)  
plt.title("Histogram of Shampoo Sales")  
plt.xlabel("Sales")  
plt.ylabel("Frequency")  
plt.grid(True)  
plt.show()  
A graph with blue lines

AI-generated content may be incorrect.

**Interpretation:** The histogram displays the frequency distribution of shampoo sales over the observed period. The x-axis represents the range of sales values (in units), while the y-axis shows the number of months falling into each sales range. The sales are divided into 10 bins, meaning the entire sales range is broken into 10 intervals. The KDE (Kernel Density Estimation) line superimposed on the bars provides a smoothed curve showing the probability density of the data. We observe a peak in the middle bins, indicating that most of the sales values lie within the mid-range (approximately 150–300 units). The curve is slightly right-skewed, suggesting that a few higher sales values exist in the dataset, which are pulling the tail towards the right. The distribution indicates that while moderate sales are most common, there were months with very high sales that are less frequent.

*# BOX PLOT*plt.figure(figsize=(6, 4))  
sns.boxplot(x=df["Sales"], color='lightgreen')  
plt.title("Boxplot of Shampoo Sales")  
plt.xlabel("Sales")  
plt.show()  
A graph showing a green rectangular bar

AI-generated content may be incorrect.

**Interpretation:** The boxplot provides a summary of the shampoo sales distribution in terms of the five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. The box represents the interquartile range (IQR), which contains the middle 50% of the data. The line inside the box shows the median, indicating the central tendency. In this case, the median lies closer to Q1, suggesting a right-skewed distribution. The whiskers extend from the box to the minimum and maximum values within 1.5 times the IQR. Any data points beyond this range are considered outliers and are shown as individual dots. These outliers represent months with exceptionally high sales, confirming the presence of anomalies in the data. This visualization is useful for identifying variability and extreme values.

*# LINE PLOT*plt.figure(figsize=(10, 5))  
plt.plot(df["Sales"], marker='o', linestyle='-', color='orange')  
plt.title("Shampoo Sales Over Time")  
plt.xlabel("Time (Index)")  
plt.ylabel("Sales")  
plt.grid(True)  
plt.show()  
  
A graph with orange lines

AI-generated content may be incorrect.

**Interpretation:** The line plot illustrates the trend of shampoo sales over time. The x-axis represents the time index (each index corresponding to a month), and the y-axis shows the sales figures. Each point on the line corresponds to sales in a particular month. The pattern indicates a general upward trend, which suggests an overall increase in sales over the period. However, we can observe fluctuations—sharp rises and falls—indicating that sales were not steady and may have been influenced by seasonal factors, marketing campaigns, or consumer demand changes. Some months saw significant increases (sharp peaks), while others showed declines (troughs). This plot is helpful in identifying long-term trends, seasonality, and cycles in the data.

*# Z SCORES*plt.figure(figsize=(8, 4))  
sns.histplot(z\_scores, kde=True, color='salmon', bins=10)  
plt.title("Z-Score Distribution")  
plt.xlabel("Z-score")  
plt.ylabel("Frequency")  
plt.axvline(0, color='black', linestyle='--')  
plt.axvline(2, color='red', linestyle='--', label="Z = 2")  
plt.axvline(-2, color='red', linestyle='--', label="Z = -2")  
plt.legend()  
plt.show()

A graph of a bar graph

AI-generated content may be incorrect.  
**Interpretation:** The Z-score distribution graph shows the standardized values (Z-scores) of sales data. A Z-score represents how many standard deviations a data point is from the mean. The x-axis shows the Z-scores, and the y-axis shows their frequency. The KDE line overlays the bars to indicate the probability density. Most of the values lie between Z = -2 and Z = +2, which is typical for a normal distribution. Vertical lines at Z = 0, +2, and -2 are used as reference points. Z = 0 indicates the mean, and values beyond ±2 suggest outliers or unusual sales figures. The presence of values with Z-scores greater than 2 confirms that some months experienced abnormally high sales. This visualization helps to identify how spread out the data is and detect potential outliers or anomalies.