**Sales dataset results**

Sample Dataset: Sales Data for a Retail Store

| Transaction ID | Date | Product | Quantity Sold | Revenue ($) |
| --- | --- | --- | --- | --- |
| 1 | 2023-01-01 | Laptop | 5 | 5000 |
| 2 | 2023-01-02 | Smartphone | 10 | 2500 |
| 3 | 2023-01-03 | Tablet | 8 | 1600 |
| 4 | 2023-01-04 | Headphones | 20 | 400 |
| 5 | 2023-01-05 | Laptop | 4 | 4000 |
| 6 | 2023-01-06 | Smartphone | 15 | 3750 |
| 7 | 2023-01-07 | Tablet | 12 | 2400 |
| 8 | 2023-01-08 | Headphones | 18 | 360 |
| 9 | 2023-01-09 | Laptop | 6 | 6000 |
| 10 | 2023-01-10 | Smartphone | 12 | 3000 |

**Dataset:** You are provided with a sample sales dataset containing information about transactions for a retail store. The dataset includes the following columns:

Transaction ID

Date

Product

Quantity Sold

Revenue ($)

**Instructions:**

**Part 1: Descriptive Statistics**

Data Preparation

Create an Excel spreadsheet.

Enter the provided dataset into the spreadsheet with appropriate column headers.

**Descriptive Statistics**

Calculate the following descriptive statistics for the 'Quantity Sold' and 'Revenue ($)' columns:

Mean (Average)

Median (Middle Value)

Mode (Most Frequent Value)

Range (Max - Min)

Variance

Standard Deviation

**Interpretation**

Write a brief interpretation of the descriptive statistics you calculated. Discuss any insights or trends that can be observed from the data.

|  |  |  |
| --- | --- | --- |
| AVERAGE | 11 | 2901 |
| MEDIAN | 11 | 2750 |
| MODE | 12 | #N/A |
| MAX RANGE | 20 | 6000 |
| MIN RANGE | 4 | 360 |
| ST DEV | 5.456901848 | 1848.846187 |
| VARIANCE | 29.77777778 | 3418232.222 |

INTERPRETATION

Average (Mean):

The average value is 2901. This represents the central tendency of the data, indicating that, on average, the values are around 2901.

Median:

The median value is 2750. The median is another measure of central tendency and is less sensitive to extreme values than the mean. It suggests that half of the values are below 2750 and half are above.

Mode:

The mode is not available (#N/A). This could mean that there is no value that appears more frequently than others in the dataset.

Maximum Range:

The maximum range is 6000. This represents the difference between the maximum and minimum values in the dataset, indicating the spread of the data. A larger range suggests greater variability in the data.

Minimum Range:

The minimum range is 360. This is the difference between the minimum and maximum values, providing another measure of data spread.

Standard Deviation:

The standard deviation is 5.46. This measures the amount of variation or dispersion in the dataset. A lower standard deviation indicates that the data points tend to be close to the mean.

Variance:

The variance is 3418232.22. This is the square of the standard deviation and provides another measure of how spread out the values are. A higher variance indicates greater variability in the data.

Overall, the data appears to have a moderate amount of variability, as indicated by the standard deviation and variance. The absence of a mode suggests that there may not be a clear peak in the distribution. The range statistics provide insights into the spread of the data, with a difference between the maximum and minimum values.

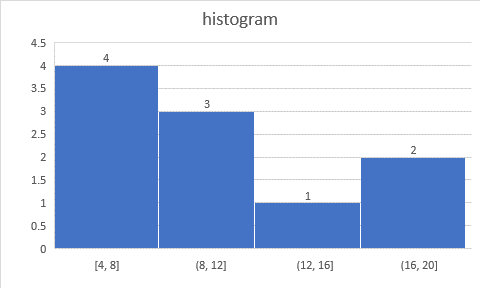
**Part 2: Data Visualization**

Histogram

Create a histogram for the 'Quantity Sold' column.

X axis =product

Y axis =quantity sold

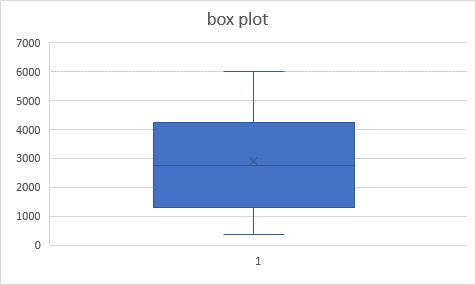


**Choose an appropriate number of bins for clarity.**

Number of bins​≈3.16 Round up to the nearest whole number: 4 bins.

**Label the axes, add a title, and provide a brief description of what the histogram illustrates.**

**Box Plot**



**Create a box plot for the 'Revenue ($)' column.**

**Label the axes, add a title, and provide a brief explanation of what the box plot reveals about revenue distribution**.

* X axis = product
* Y axis = revenue
* The minimum value is $360.
* First Quartile (Q1):
* The first quartile is the median of the lower half of the dataset. In this case, it's the median of the values $360, $1600, $2400, and $2500. The approximate value of Q1 is $2000.
* Median (Q2):
* The median is the middle value of the dataset when it is ordered. In this case, it's the fifth value when sorted, which is $3750.
* Third Quartile (Q3):
* The third quartile is the median of the upper half of the dataset. In this case, it's the median of the values $4000, $5000, $6000. The approximate value of Q3 is $5000.
* Maximum (Q3 + 1.5 \* IQR):
* The maximum value is $6000.
* Interquartile Range (IQR):
* The IQR is the range between the first quartile (Q1) and the third quartile (Q3). In this case, it's $5000 - $2000 = $3000.
* Now, let's describe how these statistics are represented in a box plot:
* Box (Interquartile Range):
* The box represents the interquartile range (IQR), spanning from Q1 to Q3. In this case, the box would extend from approximately $2000 to $5000.
* Line Inside the Box (Median):
* A line inside the box represents the median, which is at $3750.
* Whiskers (Minimum and Maximum):
* Whiskers extend from the box to the minimum and maximum values within 1.5 times the IQR. The whisker on the lower end would extend to $360, and the whisker on the upper end would extend to $6000.

**Scatter Plot**

**Create a scatter plot to visualize the relationship between 'Quantity Sold' and 'Revenue ($)'. Each point on the plot represents a transaction**.

**Scatter plot:**

Label the axes, add a title, and analyze the scatter plot's implications for understanding the relationship between quantity sold and revenue.

* X axis=quantity sold
* Y axis = revenues
* A scatter plot is a graphical representation of the relationship between two continuous variables. In this question ,we have "Quantity Sold" on the x-axis and "Revenues" on the y-axis. Each point on the scatter plot corresponds to a specific observation in dataset. Let's interpret the scatter plot based on the given data:
* **Positive Correlation:**
* The scatter plot appears to show a generally positive correlation between "Quantity Sold" and "Revenues." As the quantity sold increases, there is a tendency for the revenue to increase as well.
* **Outliers:**
* Observations such as (20, 400) and (6, 6000) may be considered outliers. These points deviate from the overall trend in the data. The point (20, 400) suggests that despite a high quantity sold (20), the revenue is relatively low.
* **Clustering:**
* Points (5, 5000), (10, 2500), and (15, 3750) seem to cluster together, indicating a consistent pattern where higher quantities sold are associated with higher revenues.
* **Non-linear Relationship:**
* The relationship between quantity sold and revenues doesn't appear perfectly linear. It seems that, particularly at higher quantities sold, the increase in revenue may not be as steep.
* **Variability:**
* There is variability in revenues for a given quantity sold. For example, at a quantity sold of 10, revenues vary between 2500 and 3000.
* Interpreting scatter plots often involves identifying patterns, trends, and potential outliers in the data. In this case, we observe the general trend of increasing revenues with increasing quantities sold, but there are variations in revenue for a given quantity sold.