1. Calculate the mean, median, standard deviation, and skewness of the variable “Quantity”. Create a Histogram and a Boxplot. Compare the mean with the median to analyze the skewness of the distribution, and provide your interpretation of the histogram and box plot.

Code:

# Load the data

sales\_data <- read\_excel("Sales\_orders.xlsx") %>%

# Calculate statistics for 'Quantity'

mean\_quantity <- mean(sales\_data$Quantity)

median\_quantity <- median(sales\_data$Quantity)

std\_quantity <- sd(sales\_data$Quantity)

skewness\_quantity <- skewness(sales\_data$Quantity)

# Print the statistics

print(paste("Mean Quantity:", mean\_quantity))

print(paste("Median Quantity:", median\_quantity))

print(paste("Standard Deviation:", std\_quantity))

print(paste("Skewness:", skewness\_quantity))

OutPut:

> print(paste("Mean Quantity:", mean\_quantity))

[1] "Mean Quantity: 3.68150346191889"

> print(paste("Median Quantity:", median\_quantity))

[1] "Median Quantity: 3"

> print(paste("Standard Deviation:", std\_quantity))

[1] "Standard Deviation: 2.20473031393098"

> print(paste("Skewness:", skewness\_quantity))

[1] "Skewness: 1.30225703471571"

# Histogram of Quantity

hist(sales\_data$Quantity, breaks = 20, main = "Histogram of Quantity", xlab = "Quantity", col = "orange")

A graph of a number of bars

Description automatically generated

**Analysis of Histogram graph:** The average quantity amount is slightly greater than the median amount. In the right-skewed distribution, means are typically greater than medians due to high-value anomalies pulling the mean to the right. The dominant categories in the histogram are two and three since most stock-keeping units have these attributes. A standard deviation of 2.20 implies that there is a moderate dispersion of values from the average value regarding quantity.

Additionally, from the histogram, it is apparent that some high values (8, 10, 12, 14) occur less frequently hence there may be the possibility of the presence of outliers or less common larger orders. Around lower quantities, sales are concentrated hence concentrating on smaller, more frequent orders might be considered.

The median quantity is shown by the thick horizontal line in the purple box is 3. Thus, half of the orders are three or less in quantity, while the other half have more than three items. When sales are concentrated on lower quantities, it might be wise to concentrate on smaller, more numerous orders as well. This may mean that a few customers are buying in bulk, or there might be some sales promotions going on.

# Boxplot of Quantity

boxplot(sales\_data$Quantity, main = "Boxplot of Quantity", ylab = "Quantity", col = "purple")

A diagram of a box plot

Description automatically generated

The median quantity, as shown by the thick horizontal line in the purple box, is 3. Thus, half of the orders are three or less in quantity, while the other half have more than three items. When sales are concentrated on lower quantities, it might be wise to concentrate on smaller, more numerous orders as well. This may mean that a few customers are buying in bulk, or there might be some sales promotions going on.

1. Create a box plot of “Quantity” by subcategory of products (variable “Sub-category”). Use subcategory in the x axis, and the box plot of “Quantity” for each “Subcategory” in the y axis to allow comparisons. Provide your interpretation and comparison analysis.

# Create Box Plot for 'Quantity' by 'Sub-category'›‹

ggplot(sales\_data, aes(x = `Sub-Category`, y = Quantity)) +

geom\_boxplot(aes(fill = 'Sub-Category')) +

theme(axis.text.x = element\_text(angle = 50, hjust = 1))

labs(title = "Box Plot of Quantity by Subcategory", x = "Subcategory", y = "Quantity")

A graph with different colored bars

Description automatically generated

"There are multiple types of subcategories, among which there are outliers such as 'Binders', 'Chairs', 'Furnishings', and 'Tables'. These isolated cases in turn suggest customers occasionally make bigger orders in those areas. Several category types have more variation in them compared to others where there might not be much dispersion or areas that do not contain any outliers.

The 'Accessories,' 'Labels,' and 'Supplies' subcategories have smaller whiskers and a decrease in outlier values concerning other such categories. This may mean there is less randomness in the number of orders made. When orders vary widely within certain product categories, one needs to adapt to more frequent changes per product rather than keeping it the same all the time. On the other hand, these constant ones should have constant stock levels for stability.

1. Analyze the mean, median, standard deviation of the variable “Profit”. Use subcategory in the x axis, and the box plot of “Profit” for each “Subcategory” in the y axis to allow comparisons. Provide your interpretation and comparison analysis.

#3.1

profit\_summary <- sales\_data %>%

group\_by(`Sub-Category`) %>%

summarise(

Mean\_Profit = mean(Profit),

Median\_Profit = median(Profit),

SD\_Profit = sd(Profit)

)

# Print summary statistics

print(profit\_summary)

**OUTPUT:**

`Sub-Category` Mean\_Profit Median\_Profit SD\_Profit

<chr> <dbl> <dbl> <dbl>

1 Accessories 22.9 4.65 85.1

2 Appliances 27.9 11.5 55.9

3 Art 3.27 1.44 4.43

4 Binders -4.83 -2.64 132.

5 Bookcases -55.6 -27.1 118.

6 Chairs -44.8 -19.3 89.1

7 Copiers 529. 180. 994.

8 Envelopes 23.9 16.4 28.9

9 Fasteners 3.57 1.87 4.99

10 Furnishings 7.37 5.78 20.9

11 Labels 9.93 7.50 9.66

12 Machines -933. -338. 1700.

13 Paper 13.3 6.04 22.9

14 Phones -58.6 -28.8 83.2

15 Storage -19.5 -2.44 75.2

16 Supplies -115. -2.19 313.

17 Tables -138. -101. 148.

#3.2

ggplot(sales\_data, aes(x = `Sub-Category`, y = Profit)) +

geom\_boxplot(fill = "orange", color = "black", alpha = 1) +

ggtitle("Box Plot of Profit by SubCategory") +

xlab("Sub-Category") +

ylab("Profit") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 50, hjust = 1))

A graph with lines and dots

Description automatically generated

Subcategories like 'Copiers' and 'Machines' have a higher median profit, while 'Tables', 'Supplies', and 'Furnishings' show negative median profits, indicating losses. Subcategories such as 'Copiers' and 'Machines' exhibit higher variability, as shown by the length of their whiskers and the presence of outliers.

Subcategories with high median profits and low variability, such as 'Copiers' and 'Machines', are likely consistent profit centers. Subcategories with negative median profits or high variability might need attention to identify and address the causes of losses or instability.