Report 1

Sana Zakir Shaikh

2025-07-24

# 📦 Data Cleaning

In this section, we load the dataset, check for missing values, clean column names, and create a combined date column.

```{r setup, include=TRUE} library(readxl) library(janitor) library(dplyr) library(lubridate)

Load data

data <- read\_excel(“mobile.xlsx”)

Clean column names

data <- clean\_names(data)

Check for missing values

sum(is.na(data)) colMeans(is.na(data)) \* 100

Create full date column

data <- data %>% mutate( full\_date = make\_date(year = year, month = month, day = day) )

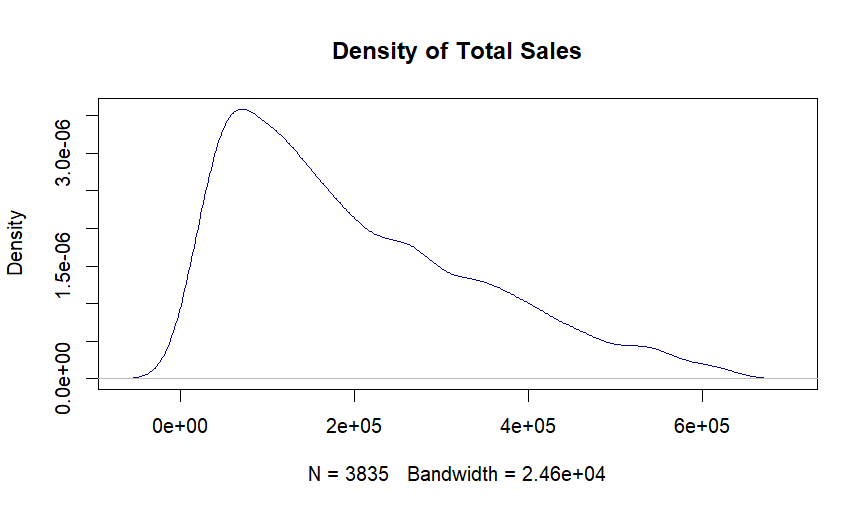
# Calculate total sales

dataunits\_sold \* data$price\_per\_unit

# Plot histogram and density

hist(data$total\_sales, main = “Histogram of Total Sales”, xlab = “Total Sales”, col = “skyblue”, breaks = 40)

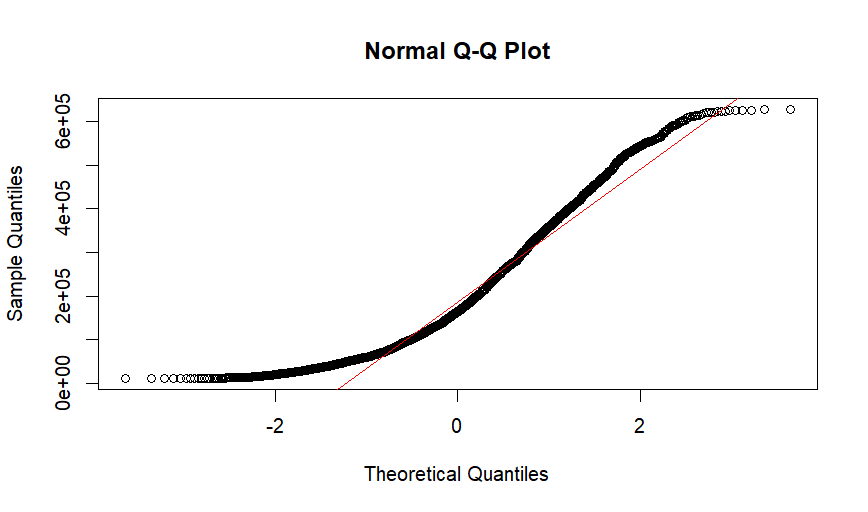
plot(density(data$total\_sales), main = “Density Plot of Total Sales”, xlab = “Total Sales”, col = “darkgreen”, lwd = 2)



Density Plot - Raw

# QQ plot

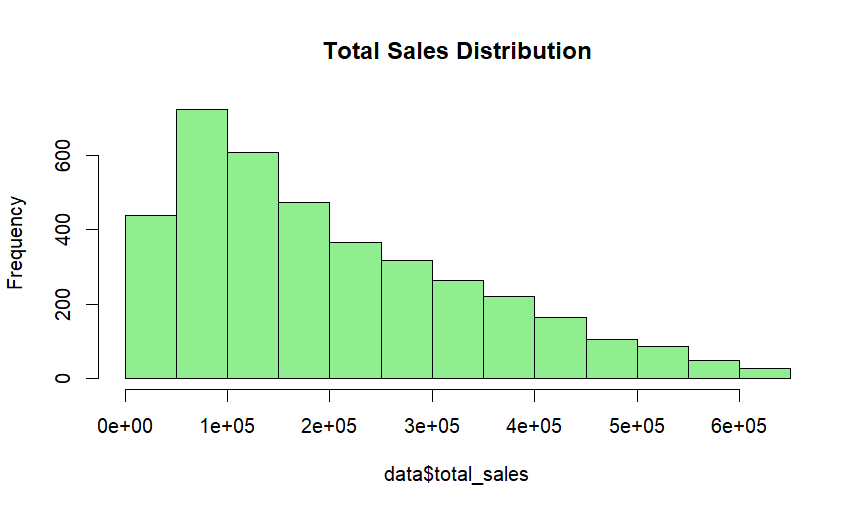
qqnorm(data$total\_sales, main = "QQ Plot of Total Sales")
qqline(data$total\_sales, col = “red”)



QQ Plot - Raw

datatotal\_sales)

plot(density(data$log\_sales), main = “Density Plot After Log Transformation”, xlab = “Log(Total Sales)”, col = “blue”, lwd = 2)



Density After Log Transformation (Left Skewed)

datatotal\_sales) plot(density(data$sqrt\_sales), main = “Density After Square Root Transformation”, xlab = “Sqrt(Total Sales)”, col = “purple”, lwd = 2)

## Comparison of Total Sales Transformations

We applied transformations to the Total Sales column to normalize the distribution. Below is a comparison between the original, log-transformed (left-skewed), and sqrt-transformed (approximately normal) distributions.

```{r comparison-density-plots, fig.width=12, fig.height=4, echo=TRUE, message=FALSE, warning=FALSE} # Required Libraries library(ggplot2) library(patchwork)

# Create the three plots

p1 <- ggplot(data, aes(x = total\_sales)) + geom\_density(fill = “darkgreen”, alpha = 0.6) + labs(title = “Original (Right Skewed)”, x = “Total Sales”, y = “Density”) + theme\_minimal()

p2 <- ggplot(data, aes(x = log\_sales)) + geom\_density(fill = “blue”, alpha = 0.6) + labs(title = “Log Transformed (Left Skewed)”, x = “Log(Total Sales)”, y = “Density”) + theme\_minimal()

p3 <- ggplot(data, aes(x = sqrt\_sales)) + geom\_density(fill = “purple”, alpha = 0.6) + labs(title = “Sqrt Transformed (Nearly Normal)”, x = “Sqrt(Total Sales)”, y = “Density”) + theme\_minimal()

# Combine plots side-by-side

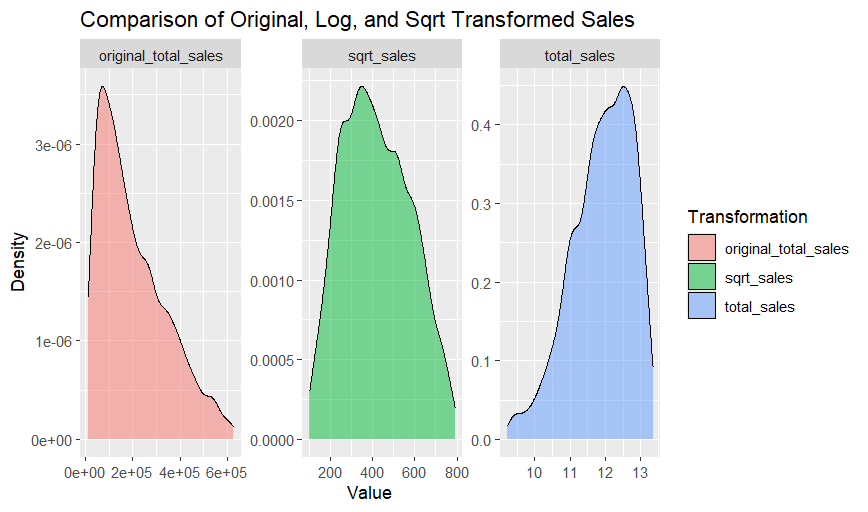
p1 + p2 + p3

## Comparison of Distribution Shapes

We observed how different transformations affect the shape of the Total Sales distribution.

Below is the image showing:

* Original (right skewed)
* Log transformed (left skewed)
* Square root transformed (approximately normal)



Final Comparison of Density Plots