1. What I have done:

I performed several tasks in exploratory data analysis part using R, focusing on: checking for outliers, skewness, and normality.

1. **Checked for Outliers:**
   - I started by identifying potential outliers in the dataset. I used boxplots to visually identify extreme values that fall outside the interquartile range (IQR). These values were marked as outliers if they were outside the upper and lower whiskers of the boxplot (i.e., values beyond 1.5 times the IQR above the 75th percentile or below the 25th percentile).
   - I also calculated the Z-scores for numerical variables to flag values that deviate significantly from the mean. A Z-score above 3 or below -3 typically indicates an outlier.
2. **Checked for Skewness:**
   - I calculated the skewness for each numerical variable in the dataset. Skewness indicates whether the data is symmetrically distributed or if it has a long tail on one side.
   - Positive skewness indicates a distribution where the right tail is longer (right-skewed), and negative skewness indicates a left-skewed distribution.
   - I used the skewness() function from the e1071 package to calculate the skewness for each column.
3. **Checked for Normality:**
   - I performed normality tests on the dataset using the **Shapiro-Wilk test** to determine if the data follows a normal distribution.
   - The **Shapiro-Wilk test** is commonly used for checking normality, and if the p-value is less than 0.05, it indicates that the data is significantly different from a normal distribution.
   - I also created **Q-Q plots** (quantile-quantile plots) for visual inspection of normality. A Q-Q plot compares the quantiles of the data with the quantiles of a normal distribution. If the data points fall along a straight line, the data is normally distributed.
   - In addition, I calculated the **kurtosis** (measuring the "tailedness" of the distribution) to further evaluate the data's departure from normality.

2. What methods I used:

1. **For Outliers:**
   - I used **boxplots** to identify outliers visually.
   - I calculated **Z-scores** for each numerical variable using the formula:

$$Z = \frac{X - \mu}{\sigma}$$

where X is the data point, μ is the mean of the variable, and σis the standard deviation. Any Z-score greater than 3 or less than -3 was flagged as an outlier.

  - I applied the **IQR method** to detect outliers. The IQR is calculated as the difference between the 75th percentile (Q3) and the 25th percentile (Q1). Data points beyond the range of Q1−1.5×IQR and Q3+1.5×IQR are considered outliers.

2. **For Skewness:**
   - I used the skewness() function from the **e1071** package to calculate the skewness for each variable. The function returns a value:
     - **Skewness > 0** indicates a right-skewed distribution (long tail to the right).
     - **Skewness < 0** indicates a left-skewed distribution (long tail to the left).
     - **Skewness ≈ 0** indicates a symmetric distribution.
   - I calculated the skewness for the variables Price, Volume, and Returns to check how their distributions deviate from symmetry.
3. **For Normality:**
   - I performed the **Shapiro-Wilk test** using the shapiro.test() function. This test checks if the sample data comes from a normally distributed population. The null hypothesis is that the data is normally distributed, and a p-value less than 0.05 indicates that the data is not normally distributed.
   - I used **Q-Q plots** to visually assess normality. I created these plots using the qqnorm() and qqline() functions, which help in visually comparing the distribution of the data to a normal distribution.
   - I also calculated **kurtosis** using the kurtosis() function from the e1071 package to evaluate the distribution's peakedness and tail behavior.

3. What results you received using the method over the data

Outlier Detection:

- **Price Data (Open/Close/High/Low)**:
  - The price data has a **value range** of approximately **0-150**, with the **median** around the **0-50** range.
  - There are **multiple outliers** concentrated in the range of **130-150**, indicating a few extreme price points that are far from the rest of the data. These outliers suggest that in some instances, the price spiked significantly higher than the typical range.
- **Volume Data**:
  - The **volume data** mainly ranges from **0 to 50,000**, with significant outliers reaching close to **250,000**.
  - There is a **long-tail distribution**, with many scattered outliers, indicating that there were numerous instances where trading volume experienced **extreme spikes**.
  - These outliers are spread across the range, implying **high volatility** in transaction volume.

Skewness Analysis:

- The **price data** (Open/Close/High/Low) shows a **skewness value of approximately 1.1**, indicating a **moderate positive skew**, which means the data has a **long right tail**, where higher values are more spread out.
- **Volume data** has a **skewness of 2.94**, suggesting a **severe positive skew**, with a concentration of low volume values and a few very large spikes at the high end. This is typical in financial data, where a majority of trading volumes are small, but some instances show extremely high values due to large transactions or market events.

Normality Testing (K-S Test & Box-Cox Transformation):

- The **Kolmogorov-Smirnov (K-S) test** for normality shows a **p-value of 0**, **strongly rejecting** the null hypothesis that the data follows a normal distribution. This result is consistent for both price and volume data, confirming that the data does not follow a normal distribution.
- After applying the **Box-Cox transformation**, the data still **did not conform** to a normal distribution, indicating that even after transformation, the data's skewness and kurtosis are too pronounced to achieve normality.
- **Optimal lambda values** for the Box-Cox transformation:
  - For **price data**: the optimal lambda is **-0.182**.
  - For **volume data**: the optimal lambda is **0.222**.

4. What is your final inference from all that you did

Key Findings:

- **Severe Skewness**: The data, both for **price** and **volume**, exhibit significant **positive skewness**. This means that both price and volume distributions have **long right tails**, with most values concentrated at the lower end of the range but a few extremely high values driving the distribution's skew.
  - For **price data**, the skewness is moderate (**1.1**), indicating that there are some price points far higher than the median, but it's not as extreme as the volume data.
  - For **volume data**, the skewness is **severe (2.94)**, suggesting that the distribution is highly skewed with frequent small values and a few extremely large spikes.
- **Non-Normal Distribution**:
  - The **K-S test results** strongly reject the hypothesis that the data follows a normal distribution (p-value = 0). This is a common feature in financial market data, where the presence of extreme events (such as market crashes or sudden spikes in trading activity) creates a **non-normal distribution**.
  - Even after applying **Box-Cox transformations**, the data **remains non-normal**, reinforcing the notion that financial data, especially prices and volumes, do not adhere to the assumptions of normality.
- **Volatility in Volume**: The **volume data** shows significant **volatility**, with many extreme spikes in the upper range. This could indicate sudden surges in trading activity, possibly due to large transactions or events that caused the market to react quickly.

Conclusion and Recommendations:

- The data has **characteristics typical of financial markets**, such as:
    - **Price data** exhibits moderate **positive skew**, which means prices are more likely to have small fluctuations but occasionally spike higher.
    - **Volume data** shows **severe positive skew**, with the possibility of large **outliers** indicating **market anomalies** or **events that caused extreme trading volumes**.
- Given the **non-normality** of the data and the presence of extreme outliers, it is recommended to use **non-parametric statistical methods** or other techniques designed for **skewed distributions**, such as **log transformations** or **rank-based methods**.
- The **extreme volume spikes** suggest that the market behaves in a volatile and unpredictable way, so **robust statistical techniques** are needed to handle such data.

Final Recommendation:

- In future analysis, I would suggest **focusing on non-parametric methods**, as they do not assume normality and are more suitable for dealing with **skewed data** and **outliers**.
- Additionally, using **robust regression models** or models that account for **volatility clustering** (such as **GARCH models**) could be useful when working with such financial data.