1. **Data Import and Preparation**

```
# Import pandas library
import pandas as pd

# Specify file path
file_path = 'C:\\Users\\loydt\\Downloads\\Projects\\Superstore Sales Dataset.xlsx'

# Read excel file into a pandas DataFrame
data = pd.read_excel(file_path)
```

   - We load the Superstore Sales Dataset using `pd.read_excel()`, storing it in a DataFrame named `data`.

2. **Filtering for High-Sales States**

```
# Define list of states of interest
states_of_interest = ['Washington', 'California', 'New York', 'Florida', 'Pennsylvania']

# Filter DataFrame to include only rows for the specified states
state_data = data[data['State'].isin(states_of_interest)]
```

   - To focus the analysis on high-sales states, we filter for specific states and create a subset, `state_data`, to represent areas with high sales volume.

3. **Log Transformation of Sales Data**

```
# Apply a lambda function to create a new column 'log_sales' containing the natural logarithm of 'Sales'
high_sales_states['log_sales'] = high_sales_states['Sales'].apply(lambda x: math.log(x) if x > 0 else None)
```

   - Since the sales data may have high variance or skewness, we apply a log transformation, which reduces skewness and stabilizes variances across categories, preparing the data for ANOVA. Using `log_sales` helps achieve a more normal distribution, improving the accuracy of statistical tests like ANOVA, which assume data normality.

4. **Boxplot Visualization for Log-Transformed Sales**

```
# Iterate through each column in 'columns_to_plot'
for column in columns_to_plot:
  # Create a boxplot figure using plotly express
  boxplot_fig = px.box(high_sales_states, x=column, y="log_sales", title=f"Log-Transformed Sales by {column}", ...)
```

   - Boxplots are generated for each categorical variable. Boxplots visualize the distribution of sales across categories, highlighting median values, interquartile ranges, and potential outliers. This visual check helps identify

whether variances appear consistent, aiding in deciding if an ANOVA test is appropriate.

Statistical Tests for Homogeneity of Variance

Before performing ANOVA, it's essential to verify the assumption of **homogeneity of variances** (similar variability across groups), which is critical for reliable ANOVA results. We use Levene's and Bartlett's tests here:

1. **Levene's Test**

```
# Perform Levene's test for homogeneity of variances
levene_stat, levene_p = levene(*groups)
```

   - Levene's test assesses the equality of variances across groups. It's more robust to deviations from normality than Bartlett's test, making it useful when data may not follow a perfectly normal distribution.
   - **Interpretation**: If the p-value < 0.05, we reject the null hypothesis, indicating that variances are significantly different, which could affect ANOVA's accuracy.

2. **Bartlett's Test**

```
# Perform Bartlett's test for homogeneity of variances
bartlett_stat, bartlett_p = bartlett(*groups)
```

   - Bartlett's test also checks for homogeneity of variances. However, it is sensitive to normality assumptions, so it's best used when data is approximately normal.
   - **Interpretation**: Like Levene's test, a p-value < 0.05 indicates unequal variances, suggesting a violation of ANOVA assumptions. If both tests show unequal variances, ANOVA results may be unreliable, and we might consider alternative tests (e.g., Welch's ANOVA).

Performing ANOVA Tests

3. **ANOVA Test**

```
# Perform one-way ANOVA using the 'f_oneway' function
anova_result = f_oneway(*groups)
```

   - **One-Way ANOVA** evaluates whether there are significant differences between the means of multiple groups. Here, we analyze the log-transformed sales data across categories like Sub-Category, State, Segment, Ship Mode, and Region.

- **Interpretation**: If the p-value < 0.05, we reject the null hypothesis, indicating that at least one group's mean significantly differs from the others. This result helps pinpoint specific categorical variables that have a statistically significant impact on sales.

Importance of These Methods and Visualizations

- **Log Transformation** stabilizes variances, improving the validity of statistical tests by addressing potential skewness in the data.
- **Boxplots** offer a quick visual check for outliers and general variance patterns, aiding in initial assumptions about group variances and normality.
- **Levene's and Bartlett's Tests** help verify ANOVA's homogeneity of variance assumption, ensuring that variances across groups are not significantly different.
- **ANOVA Test** identifies whether categorical variables (like state or segment) are statistically related to sales. This insight can guide business strategies by focusing on groups that show significant differences, allowing a data-driven approach to marketing, stock management, or customer segmentation.

Overall, these methods combined offer a structured approach to ensure ANOVA's assumptions are met, leading to more accurate and meaningful insights into sales trends across categorical variables.