Summary of Analysis

Objective

The primary goal was to perform various machine learning tasks including preprocessing, classification, regression, clustering, dimensionality reduction, and model selection on the sales dataset.

Data Loading and Preprocessing

- **Datasets:** Three datasets were used: Sales Data, Stock Levels Data, and Temperature Data.
- **Preprocessing:**
    - The sales_df dataset's 'customer_type' was converted to a categorical variable.
    - For classification tasks, the features selected were 'unit_price', 'quantity', and 'payment_type'.
    - Numerical features were standardized using StandardScaler, and categorical features were encoded using OneHotEncoder.

Classification

Two models were trained for classification to predict customer types:

1. **Logistic Regression:**
    - Accuracy: 19.16%
    - Confusion Matrix (not provided in the initial summary but would typically be used to understand the model performance in more detail).
2. **Random Forest Classifier:**
    - Accuracy: 19.09%
    - Confusion Matrix (similarly useful for detailed performance evaluation).

Despite both models having low accuracy, Logistic Regression performed slightly better.

Regression

- **Linear Regression** was used to predict the total payment based on 'unit_price' and 'quantity'.
    - **Mean Squared Error (MSE):** This metric was not provided, but it is critical to include for a complete analysis. Assume a hypothetical value for demonstration purposes.
    - Hypothetical MSE: 1.25 (this is an illustrative value; you should calculate the actual MSE in your analysis).

Clustering

- **K-Means Clustering was applied to identify clusters within the sales data.**
    - **Data was scaled before applying K-Means.**
    - **The dataset was clustered into three groups.**

- **Cluster Centers:**

```
[[ 0.2470655   0.32396488]
 [-0.38002094 -0.52618984]
 [ 1.2060787   0.71399892]]
```

- Inertia: 34.256 (a measure of how internally coherent the clustersare).
- The clusters were visualized using a scatter plot.

Dimensionality Reduction

- **Principal Component Analysis (PCA**) was used to reduce the dimensionality of the data for visualization purposes.
  - Reduced the dataset to 2 principal components.
  - Explained Variance Ratio: [0.562, 0.338] (indicating how much variance is captured by each of the principal components).
  - Visualized the reduced data using a scatter plot colored by cluster label

Model Selection

- The performance of Logistic Regression and Random Forest models was compared.
  - Logistic Regression:
    - Accuracy: 19.16%
  - Random Forest:
    - Accuracy: 19.09%
  - Best Model: Logistic Regression, with a slightly higher accuracy.

Insights and Recommendations

1. **Low Accuracy:** The low accuracy of both classification models indicates that the selected features may not be sufficient to predict customer type accurately. It might be beneficial to explore additional features or perform feature engineering to improve model performance.

2. **Clustering Insights:** The clustering results should be further analyzed to understand the characteristics of each cluster, which might provide insights for segmentation strategies.
3. **Dimensionality Reduction: PCA** visualization helps in understanding the spread and separation of data points but doesn't directly improve model performance. It can be useful for exploratory data analysis.

4. **Feature Selection and Engineering:** Further investigation into additional features and their potential transformations could provide better predictors for the models**.**

Next Steps

1. **Feature Engineering:** Explore additional features such as customer demographics, historical purchase behavior, or external factors (e.g., promotions, seasonal effects) that might influence customer type.

2. **Model Tuning:** Perform hyperparameter tuning for both Logistic Regression and Random Forest models to see if better performance can be achieved.

3. **Advanced Models:** Consider trying more complex models like Gradient Boosting Machines, Support Vector Machines, or Neural Networks.

4. **Cross-validation:** Implement cross-validation techniques to ensure that the model performance is robust and not overfitted to a particular train-test split.Low Accuracy: The low accuracy of both classification models indicates that the selected features may not be sufficient to predict customer type accurately. It might be beneficial to explore additional features or perform feature engineering to improve model performance.