Explanation for the first part of code with box plots and data
From the results of your anomaly detection analysis, we can draw several insights based on the two models used: Isolation Forest (IF) and Local Outlier Factor (LOF), evaluated across different contamination rates, fold sizes, and fold counts. Here are the main observations and conclusions:

1. Model Comparison:

- **Isolation Forest** generally produced lower mean anomaly scores compared to LOF across most contamination rates and fold sizes, indicating that it is potentially more conservative with anomaly scores.
- **LOF** showed consistently higher mean anomaly scores across different configurations, suggesting it may classify more observations as anomalous or assigns higher anomaly levels to these points.

2. Effect of Contamination Rate:

- As the **contamination rate increased from 0.01 to 0.10**, both models identified more anomalies, which is expected since higher contamination rates make the models more sensitive to anomalies.
- For **contamination rates at 0.05 and 0.10**, both models detected significantly more anomalies across all folds, but the anomaly score for Isolation Forest dropped in magnitude as contamination increased. This could indicate that the model becomes more lenient or tolerant to outliers with a higher contamination assumption.

3. Impact of Fold Size:

- For both models, increasing the **fold size (from 5 to 10 to 15)** generally resulted in a decrease in anomalies detected for lower contamination rates (e.g., 0.01), likely due to the broader sampling of the data reducing variance in anomaly detection.
- For **contamination rates of 0.05 and 0.10**, larger fold sizes showed mixed results, with anomalies detected sometimes stabilizing or even decreasing slightly. This suggests that both models may reach a threshold of sensitivity where increased fold size does not increase anomaly detection rates significantly.

4. Consistency Across Folds:

- **Isolation Forest** shows high consistency across different folds within the same configurations, particularly for contamination rates of 0.01 and fold sizes of 5 and 10. This consistency indicates robustness and stability in its anomaly detection, making it a potentially reliable choice for datasets with similar distributions.
- **LOF**, on the other hand, shows more variability in anomalies detected across different folds, especially at higher contamination rates, which could indicate sensitivity to data distribution changes or fold sampling.

5. Optimal Model and Configuration for Anomaly Detection:

- For conservative anomaly detection, **Isolation Forest** with a contamination rate of 0.01 and a fold size of 10 shows a balanced detection rate with relatively low mean anomaly scores.
- **LOF** at a contamination rate of 0.05 and fold size of 15 provides higher anomaly scores, which could be useful for cases where more aggressive anomaly detection is preferred, though it may also result in some normal points being classified as anomalies.

Summary of Recommendations:

- **Use Isolation Forest with lower contamination rates (0.01 to 0.05) and moderate fold sizes (10 to 15)** for more conservative anomaly detection, especially when data consistency is required across different samples.
- **Apply LOF with a contamination rate of 0.05 or higher** for datasets where catching as many anomalies as possible is important, but keep in mind the potential for some variability across folds.
- Future analysis could focus on using additional metrics such as AUC or F1 score to further assess the effectiveness of each model's anomaly predictions.

Explanation for code for the second part with no of anomalies and different type of plot.

From the summarized results, we can draw several conclusions regarding the performance of the Isolation Forest and LOF (Local Outlier Factor) models for different contamination rates and fold sizes. Below are the key insights and implications of these results.

1. Effect of Contamination Rate on Mean Anomaly Scores

- **Isolation Forest**:
    - As the contamination rate increases from 0.01 to 0.10, the mean anomaly scores decrease significantly.
    - With a contamination rate of 0.01, the scores are consistently higher, ranging between 0.2461 and 0.2489, indicating that at this low contamination rate, the model identifies fewer data points as anomalies.
    - With a contamination rate of 0.10, the mean anomaly scores are around 0.1043 to 0.1067. This indicates a higher confidence in anomalies at the 0.10 contamination rate, as more points are assumed to be anomalies.
    - **Implication**: Higher contamination rates lead Isolation Forest to detect more anomalies with lower anomaly scores, indicating increased sensitivity to potential outliers in the data.
- **LOF**:
    - The LOF model shows a similar trend as Isolation Forest with decreasing mean anomaly scores as contamination rate increases.
    - At a contamination rate of 0.01, mean scores are around 0.646, indicating fewer data points as outliers with relatively high anomaly scores.
    - At a contamination rate of 0.10, scores drop to around 0.117, showing more points being identified as outliers.
    - **Implication**: The LOF model becomes more inclusive of potential anomalies at higher contamination rates, which is consistent with expectations.

2. Effect of Fold Size on Mean Anomaly Scores

- Across all contamination rates, neither model demonstrates a significant impact of fold size on the mean anomaly score.
- **Isolation Forest**:
    - Regardless of fold size (5, 10, or 15), the mean anomaly scores remain very close within each contamination rate. For example, at 0.01 contamination rate, scores are around 0.246-0.249 across fold sizes.
- **LOF**:
    - Similarly, LOF scores remain fairly consistent across fold sizes within each contamination rate. For example, for a contamination rate of 0.05, mean scores are between 0.236 and 0.239 across fold sizes.

- **Implication**: Fold size does not drastically alter the models' performance on detecting anomalies. This suggests stability in both models' sensitivity to anomalies regardless of how the data is split into folds.

3. Comparing Isolation Forest and LOF Performance

- **Mean Anomaly Scores**:
  - At each contamination level, LOF consistently produces higher mean anomaly scores than Isolation Forest.
  - At a contamination rate of 0.01, LOF scores are around 0.646, while Isolation Forest scores are around 0.246. This indicates that LOF has a higher anomaly threshold for detecting outliers compared to Isolation Forest.
- **Sensitivity to Contamination Rate**:
  - Both models exhibit decreased mean anomaly scores as contamination rate increases, but LOF's drop is more dramatic.
  - **Implication**: Isolation Forest may be more conservative and consistent in anomaly detection, while LOF adapts more aggressively as the contamination rate increases.

4. Practical Considerations

- For scenarios where a higher anomaly detection threshold is preferred, such as in datasets where anomalies are rare but impactful, **Isolation Forest with a lower contamination rate** (0.01) may be more appropriate due to its conservative approach.
- If the use case requires more aggressive anomaly detection, especially where the anomaly proportion is suspected to be higher, **LOF with a higher contamination rate** (0.05 or 0.10) could be more suitable.

Summary Points

1. **Lower contamination rates result in higher mean anomaly scores** for both models, implying fewer anomalies detected and a higher anomaly threshold.
2. **Isolation Forest is more conservative** in detecting anomalies, with lower variability in scores across fold sizes and contamination rates, while **LOF is more responsive to contamination adjustments**.
3. **Fold size has minimal impact** on both models' performance, suggesting that anomaly detection stability is achieved across different fold settings.
4. **Model selection should align with the desired sensitivity** to outliers, with Isolation Forest suitable for scenarios needing cautious anomaly detection, and LOF for more aggressive detection needs.

These insights can guide the choice and configuration of anomaly detection models based on specific project requirements and the expected anomaly frequency in the dataset.