Opis wizualizacji i wykrywania anomalii w języku polskim

1. Dendrogram metodą ward

Co przedstawia:

- **Dendrogram** pokazuje hierarchiczną strukturę klasteryzacji.
- Osie:
  - **X**: obiekty lub ich grupy po klasteryzacji.
  - **Y**: odległości między grupami (w przestrzeni cech).
- **Szczegóły:**
  - truncate_mode='level', p=6: przedstawia tylko najwyższe 6 poziomów klasteryzacji.
  - Odległości między węzłami wskazują, jak bardzo różnią się obiekty lub klastry.

Interpretacja:

- Jeśli dwa klastry łączą się na dużej wysokości, są mniej podobne.
- Użytkownik może wybrać poziom odcięcia (np. 5 klastrów), aby podzielić dane na grupy.

2. Rozkład wartości w kolumnie sub-region

Co przedstawia:

- **Wykres słupkowy** pokazuje liczbę obiektów w każdym regionie (sub-region).

Interpretacja:

- Można zobaczyć, które regiony zawierają najwięcej/najmniej obiektów w zestawie danych.
- Przydatne do analizy nierównowagi między kategoriami.

3. Wizualizacja klastrów w przestrzeni PCA

Co przedstawia:

- Klastry są przedstawione w przestrzeni dwuwymiarowej po redukcji wymiarowości za pomocą PCA.
- **Kolor** wskazuje, do którego klastra należy obiekt.

Interpretacja:

- Wizualizacja pomaga zobaczyć, jak obiekty są rozmieszczone między klastrami.
- Oddalone punkty mogą być potencjalnymi anomaliami.

Kod do wykrywania i wizualizacji anomalii

1. Wykrywanie anomalii za pomocą LOF (Local Outlier Factor):

- Algorytm LOF identyfikuje lokalne anomalie, porównując gęstość każdego punktu z jego sąsiadami.
- Punkty oznaczone jako anomalie (etykieta -1) są potencjalnymi odstępstwami od normy.

2. Wizualizacja anomalii w przestrzeni PCA:

- Na wykresie PCA zaznaczono anomalie jako czerwone punkty, co pozwala łatwo zidentyfikować, które obiekty różnią się od pozostałych.

3. Wykrywanie anomalii na podstawie odległości od centroidów:

- Algorytm mierzy odległości każdego punktu od najbliższego centroidu klastra.
- Punkty znajdujące się dalej niż 95. percentyl tych odległości są uznawane za anomalie.

Podsumowanie:

1. **Dendrogram** pozwala zobaczyć hierarchiczne grupowanie danych i wybrać poziom odcięcia.
2. **Rozkład wartości w** sub-region ujawnia nierównowagę między regionami.
3. **Wizualizacja klastrów PCA** pokazuje strukturę grup i potencjalne anomalie.
4. **LOF i odległości od centroidów** pozwalają wykryć anomalie na podstawie lokalnej gęstości lub odległości w przestrzeni klastrów.