Analiza i Klasteryzacja Danych Cocktaili

Twoje Imię

13 października 2024

1 Wprowadzenie

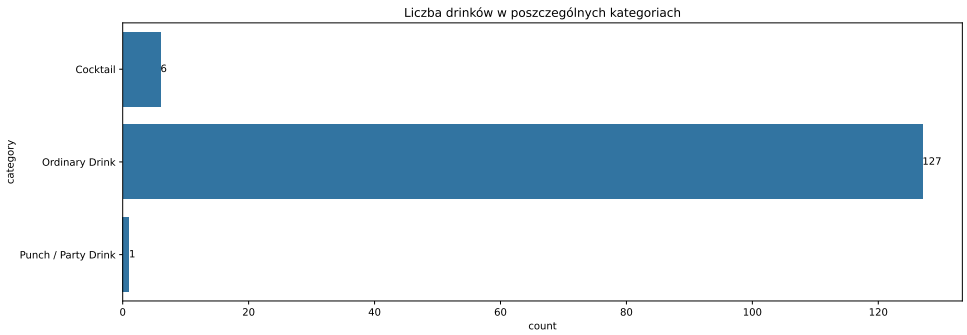
Celem tego projektu jest eksploracyjna analiza danych (EDA) oraz klasteryzacja drinków na podstawie dostępnych informacji o składnikach, kategoriach oraz typach szklanek. Projekt obejmuje przetwarzanie danych, klasteryzację oraz wizualizację wyników.

2 Zbiór Danych

Dane pochodzą z pliku `cocktail_dataset.json`, który zawiera informacje na temat różnych drinków, ich kategorii, składników oraz typów szklanek. Wstępna analiza danych obejmowała sprawdzenie struktury danych, brakujących wartości oraz rozkładu poszczególnych cech.

3 Analiza Eksploracyjna Danych (EDA)

W trakcie eksploracyjnej analizy danych zidentyfikowano rozkład poszczególnych cech, takich jak kategorie drinków i typy szklanek. Wizualizacje zostały wykonane przy użyciu pakietów `matplotlib` oraz `seaborn`. Poniżej przedstawiono przykładowe wykresy przedstawiające liczbę drinków w poszczególnych kategoriach oraz typach szklanek (Rysunki 1 i 2).



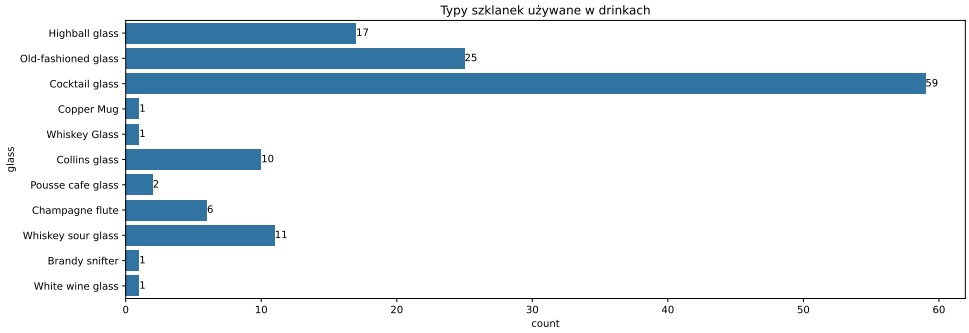
Rysunek 1: Liczba drinków w poszczególnych kategoriach

4 Preprocessing Danych

Dane zostały przetworzone, aby umożliwić ich klasteryzację. Obejmuje to czyszczenie danych, kodowanie cech kategorycznych (np. `category`) za pomocą `OneHotEncoder`, a także normalizację cech za pomocą `StandardScaler`. Proces ten zapewnia, że dane są w odpowiednim formacie do dalszej analizy.

5 Klasteryzacja

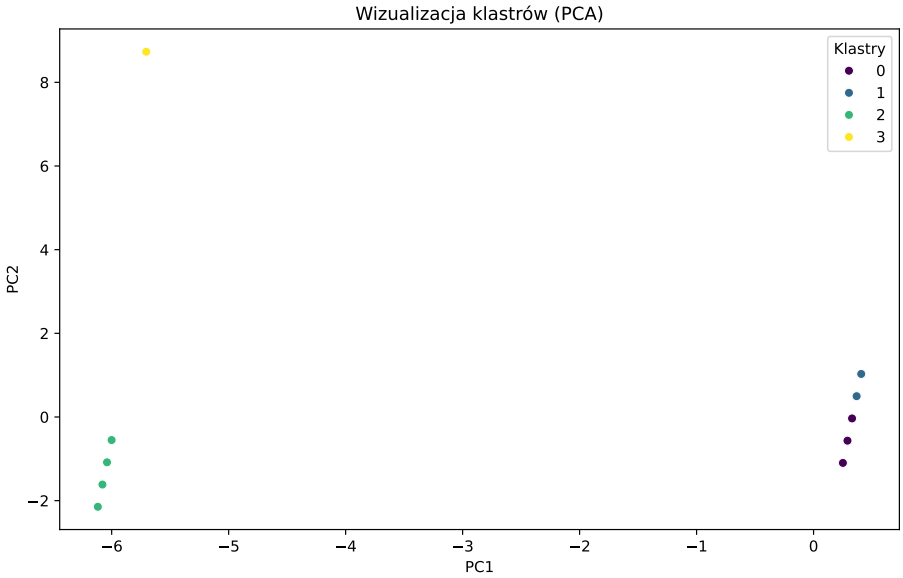
Do klasteryzacji danych użyto algorytmu `KMeans` z czterema klastrami. Model został dopasowany do danych przetworzonych, a następnie wyniki zostały ocenione za pomocą metryk, takich jak Silhouette Score oraz Davies-Bouldin Score.



Rysunek 2: Typy szklanek używane w drinkach

Silhouette Score wynosi: 0.45
Davies-Bouldin Score wynosi: 1.25

Poniżej przedstawiono wizualizację klastrów po redukcji wymiarowości za pomocą `PCA` (Rysunek 3).



Rysunek 3: Wizualizacja klastrów po redukcji wymiarowości (PCA)

6 Podsumowanie

Projekt wykazał, że klasteryzacja drinków na podstawie składników, kategorii oraz typów szklanek pozwala na wyodrębnienie grup drinków o podobnych cechach. Użyte metryki wskazują na poprawność klasteryzacji, a wizualizacja wyników umożliwiła lepsze zrozumienie struktury danych.