Analiza Danych - Sprawozdanie

Yehor Voznyi

Tematem projektu jest analiza danych przy użyciu K-średnich, wyszedłem nieco poza wymagania i dodałem kilka nowych funkcji:

1. Znaczne zmiany w oryginalnym kodzie ze względu na zmiany w bazie danych i metodzie normalizacji, a większość obliczeń została wykonana przy użyciu Pandas

2. Metodę inicjalizacji K-means++:

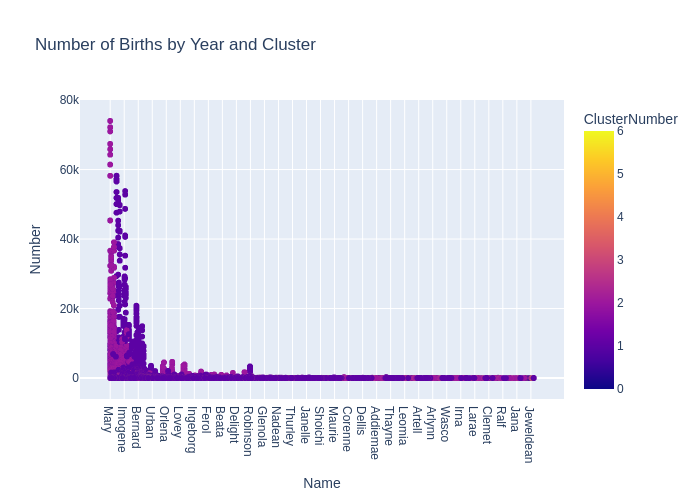
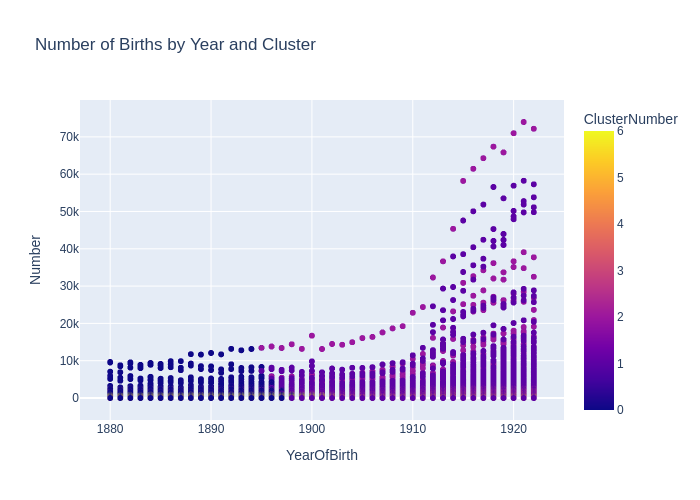
- Pierwszy centroid jest wybierany losowo.

- Dla każdego punktu danych obliczana jest minimalna odległość do już wybranych centroidów.

- Następny centroid jest wybierany na podstawie rozkładu prawdopodobieństwa proporcjonalnego do kwadratu odległości.

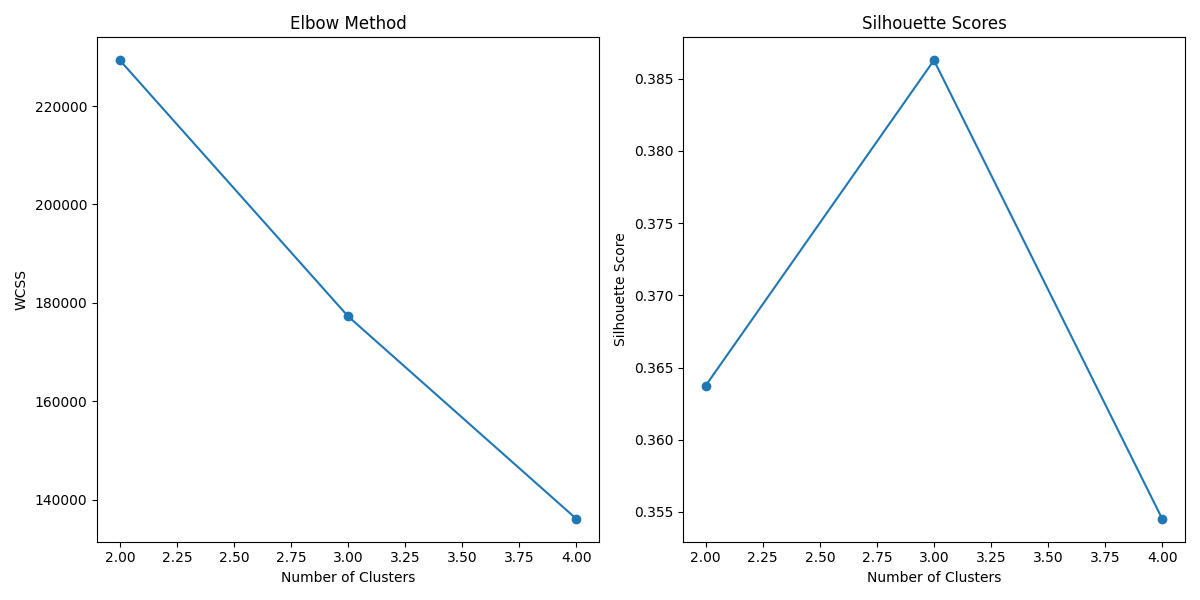
3. Automatyczna normalizacja danych: zrealizowane za pomocą bibliotek Pandas i Sklearn

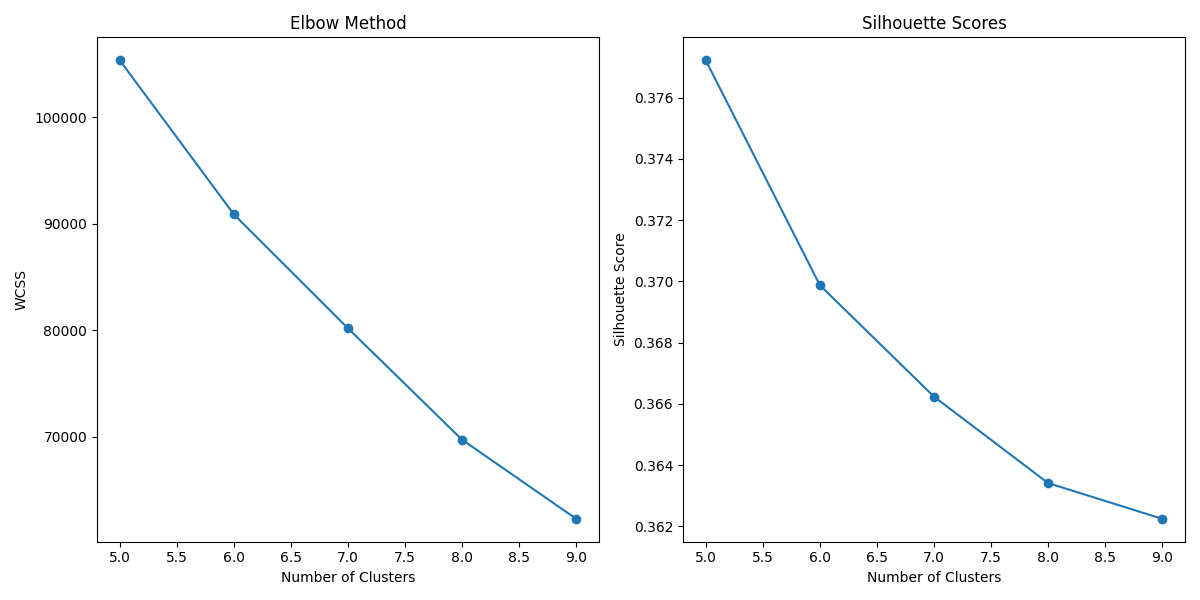
4. Wizualizacja danych w postaci wykresu rozrzutu: zrealizowane za pomocą bibliotek Pandas i Sklearn



5. Analiza danych przy użyciu dwóch metod: metody łokciowej i metoda wyników sylwetkowych za pomocą Pandas, matplotlib i sklearn

wyniki metody łokciowej i metody wyników sylwetkowych:





6. Użyłem bazy danych podobnej do oryginalnej, ale większej - dwieście tysięcy rekordów.