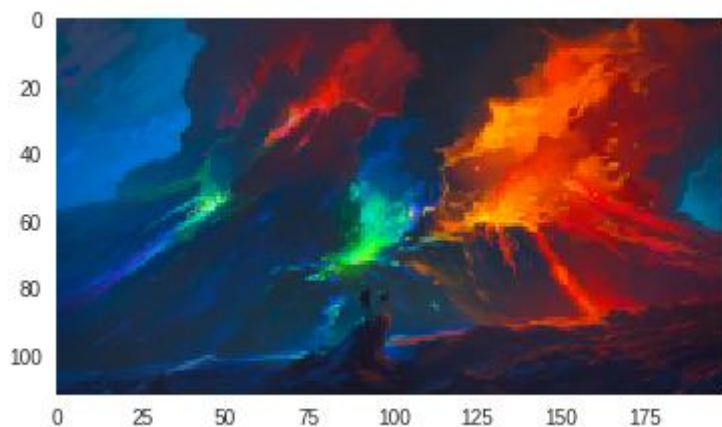
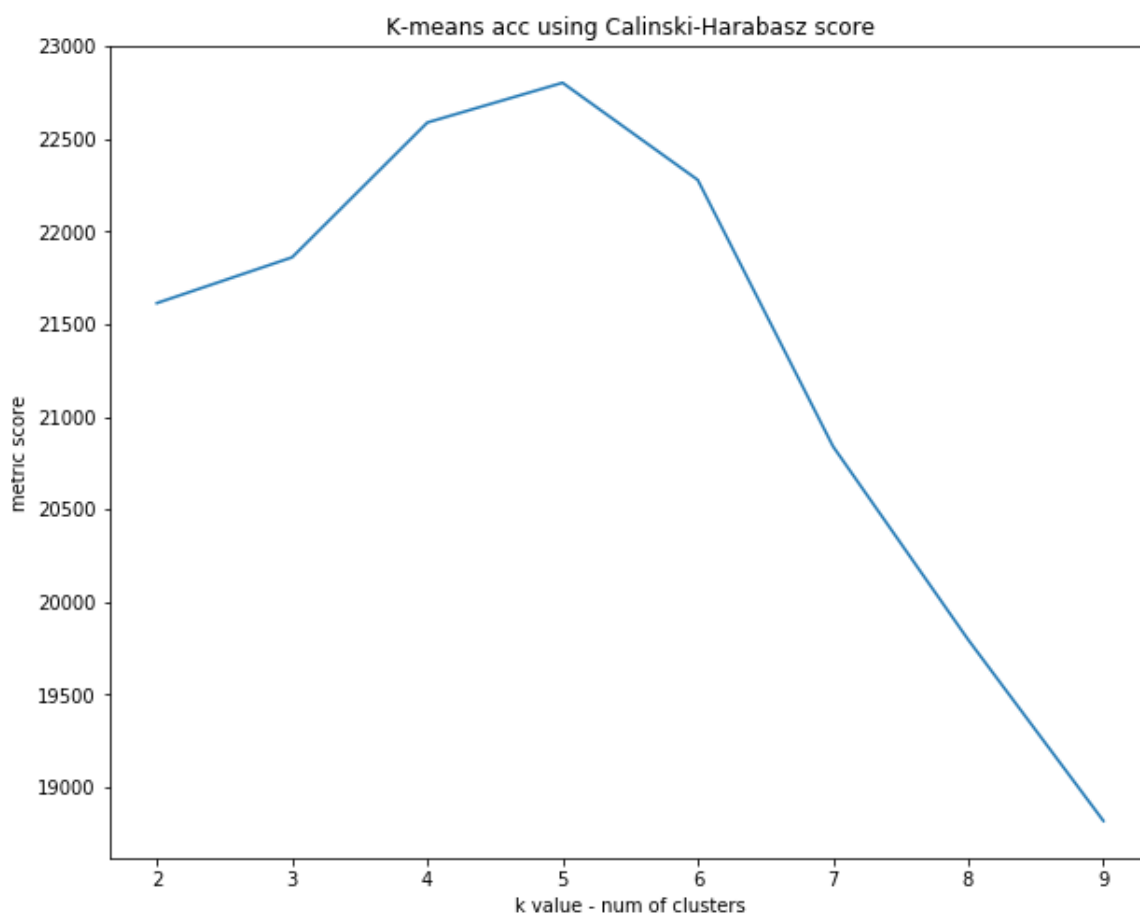


Raport – Kmean color palette
Podstawy nauczania maszynowego
Wyk. Mateusz Woś

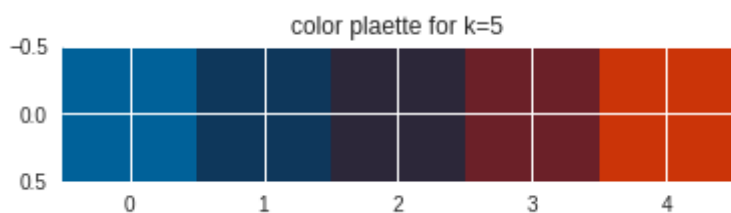
Do wykonania zadania użyłem grafiki, która zmniejszyłem do rozmiarów 200x120. Początkowo próbowałem z większymi rozmiarami, lecz skutkowało to memory error lub niesamowicie długim czasem liczenia(dla k-medoids)



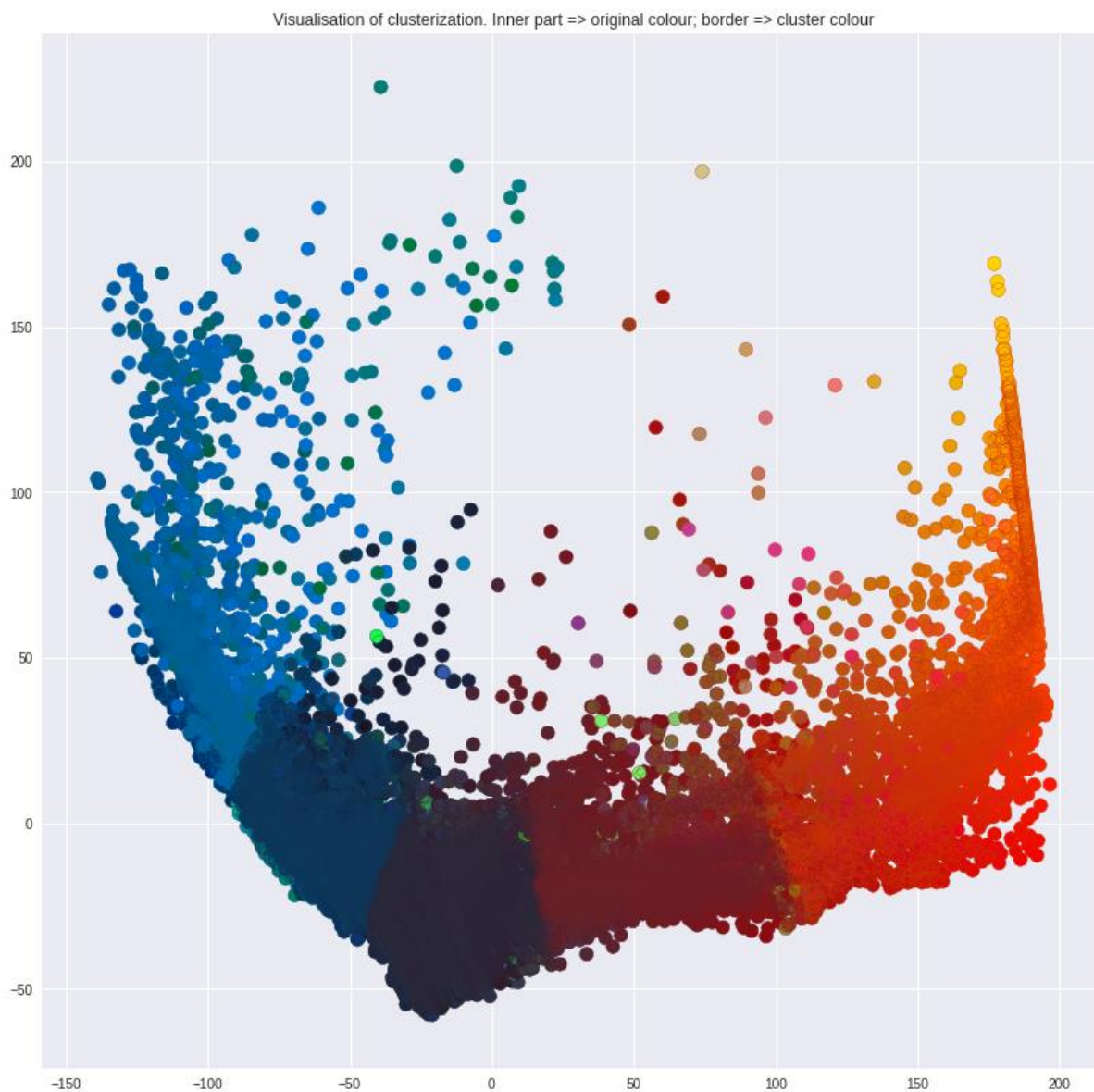
Standardowo wyeksportowałem dane do pliku csv



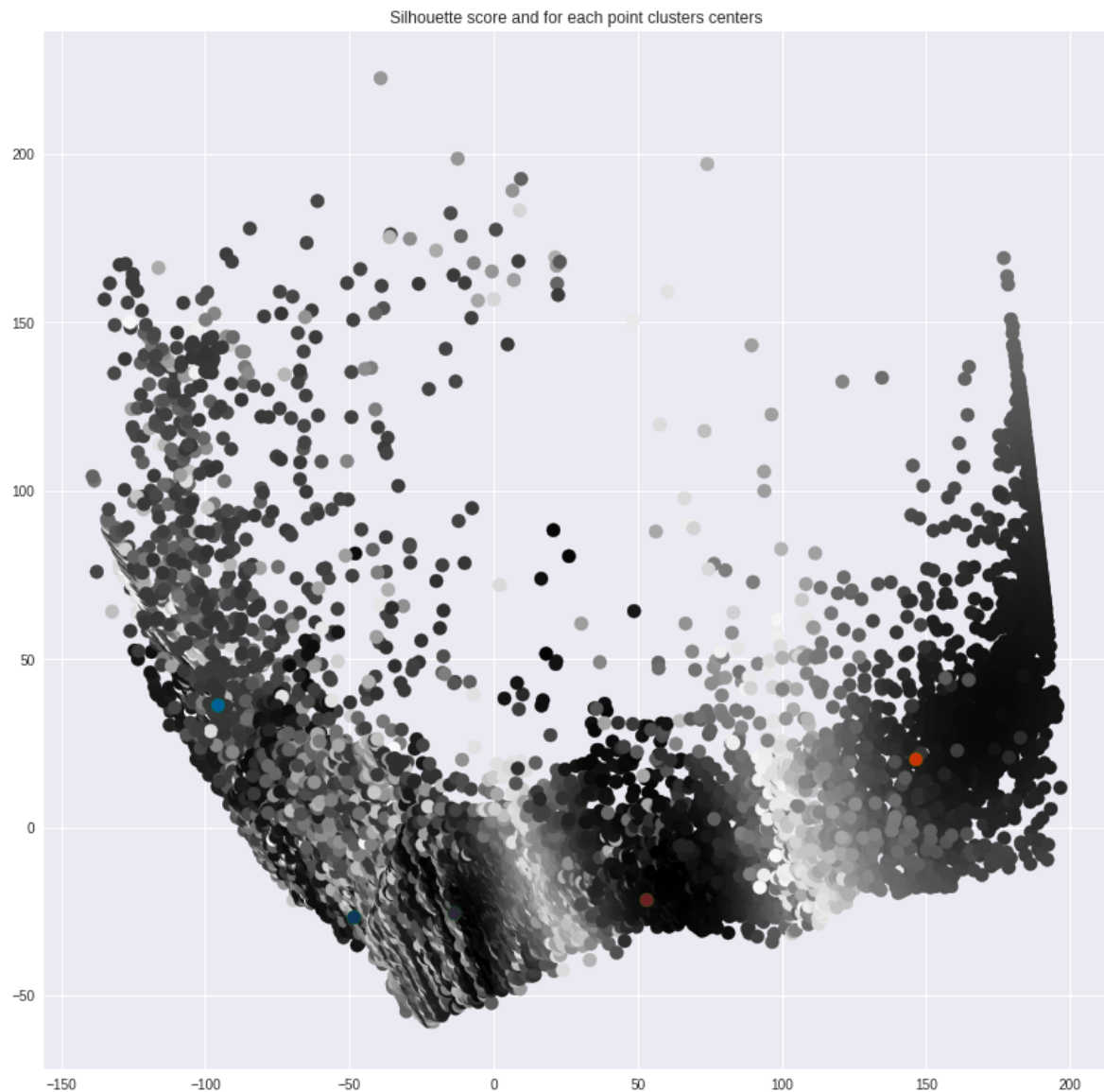
Skorzystałem z gotowej implementacji klasteryzacji k-medoids. Wyniki przedstawiłem z użyciem metryki Calinskiego-Harbasza. Dla każdego k obliczenia powtarzałem 7 krtonie. Działanie sprawdzałem dla k z zakresu od 2 do 9. Jak widac na powyzzszym wykresie, wg metryki Calinskiego-Harbasza najabrdziej optymalnym k jest k=5.



Otrzymana paleta minela się z tym czego oczekiwałem. Brak w niej koloru zielonego, pojawia się on dopiero dla k=7.



Rzut na płaszczyznie 2D pokazuje jak porozchodziły się poszczególne kolory.



Kolorami zaznaczone są centra poszczególnych klastrów. Można zaobserwować niskie wartości Silhouette score na granicach klastrów i tam gdzie punkty różnią się kolorem od klastra, w którym się znajdują.

Wnioski:

- Punkty oddalone od swoich klastrów, te na granicy klastrów i różniące się kolorami od klastrów do których zostały przyporządkowane mają niskie wartości metryki Silhouette'a. Pozwala to z pewnością wykryć anomalie w klasteryzacji.
- Wg metryki Calinskiego-Harbasza $K=5$ dla mojego datasetu jest najbardziej optymalnym k . Osobiście byłem zdziwiony, iż kolor zielony został całkowicie pominięty. Podejrzewam, że obszar ten był zbyt mały.