Po zaimportowaniu danych z pliku glass.data, sprawdzam liczbę unikatowych wartości, maksymalne i minimalne wartości dla poszczególnych kolumn(dla całości i dla poszczególnych typów szkła).

Sprawdzam też jak wyglądają zależności pomiędzy wartościami w kolumnach:Obraz zawierający zrzut ekranu, kwadrat, wzór, Prostokąt

Opis wygenerowany automatycznie

Jak widać, żadna wartość nie zależy bezpośrednio od innej wartości.

Z powodu braku informacji o szkle typu 4, podmieniam typ 7 na typ 4 żeby ułatwić przechodzenie po tablicach.

Za pomocą pandas.qcut dokonuję dyskretyzacji na kopii danych z pliku, a dzięki MinMaxScaler z sklearn.preprocessing dokonałem normalizacji kolejnej kopii danych.

Oryginalne dane, znormalizowane dane i dane na których dokonałem dyskretyzacji, łączę w jedną zmienną, dzięki czemu podział na zestaw treningowy i testowy odbędzie się identycznie dla każdego zestawu.

Zaimplementowałem dwie metody naiwnego klasyfikatora Bayesa,   
wykorzystujący średnią i medianę wartości kolumny dla danego typu szkła(który był problematyczny dla np. kategorii 6)  
wykorzystujący podział na kategorie, przy którym najlepiej radził sobie zestaw danych po dyskretyzacji.