Projekt LA

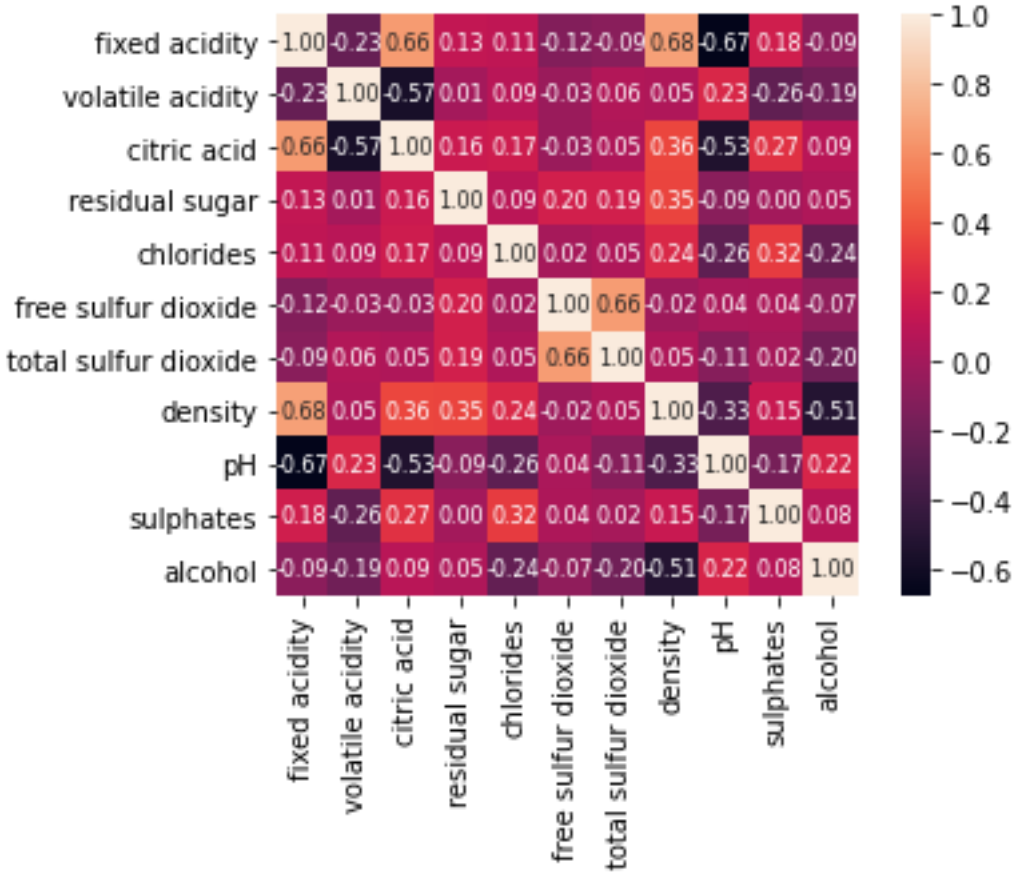
Sebastian Krzosek
300136

Cel projektu

Mając bazę danych portugalskich win czerwonych naszym zadaniem było kolejno zbudowanie modelu klasyfikacyjnego wina ze względu na ocenę jakości, modelu szacowania oceny wina oraz podzielenie win na grupy.

Przygotowanie danych

Sprawdzając poprawność danych, nie zauważyłem żadnych braków bądź niepokojących nieprawidłowości. Zwróciłem uwagę jedynie na okresowe rozwinięcia dziesiętne ułamków, więc postanowiłem zaokrąglić je do czterech miejsc po przecinku. Kolejnym krokiem było wykonanie podziału na zbiór uczący oraz testowy danych (70/30) z moim indeksem jako ziarno generatora liczb losowych. Następnie wykonałem macierz korelacji w celu sprawdzenia powiązania miedzy zmiennymi. Możemy zauważyć, że zmienne nie są ze sobą skorelowane, dlatego też wybrałem je wszystkie jako predykatory w przygotowywanym modelu. Na tym etapie, bazując jedynie na mojej wiedzy o winie, wydaje mi się, że wszystkie zmienne w podobnym stopniu wpływają na ocenę jakości wina. Żadna z nich nie wyróżnia się, natomiast po zaczerpnięciu informacji z Internetu, wiem że do klasyfikacji win na wytrawne, półwytrawne, półsłodkie i słodkie znaczenie będą miały zmienne takie jak zawartość alkoholu, cukier, kwasowość ogólna (pH) oraz kwasowość lotna.



Klasyfikacja – Algorytm KNN

Problem klasyfikacji postanowiłem rozwiązać przy użyciu algorytmu KNN. Pierwszym krokiem było przeprowadzenie standaryzacji predyktorów, a następnie uruchomienie algorytmu biorąc pod uwagę czterech najbliższych sąsiadów w metryce euklidesowej (przy tej liczbie otrzymałem najbardziej satysfakcjonujące wyniki). Ciekawym spostrzeżeniem było to, że mimo możliwości oceniania win w skali od 1 do 10, nie było wina ocenionego niżej niż 3 oraz wyżej niż 8. Po zaobserwowaniu tego stworzyłem macierz pomyłek, która posłużyła mi do policzenia trafności oraz trafności z odstępstwem o 1. Otrzymałem następujące wyniki:

```
[[  0   1   1   1   0   0]
 [  0   1  12   3   0   0]
 [  0   4 150  48   2   0]
 [  0   2  70 101  18   1]
 [  0   0   6  31  21   2]
 [  0   0   0   2   3   0]]

Trafność: 273/480 = 0.56875
```

```
[[  0   1   1   1   0   0]
 [  0   1  12   3   0   0]
 [  0   4 150  48   2   0]
 [  0   2  70 101  18   1]
 [  0   0   6  31  21   2]
 [  0   0   0   2   3   0]]

Trafność z odstepstwem o 1: 462/480 = 0.9625
```

Trafność na poziomie ≈ 57% nie jest zadowalająca jednak patrząc na duży wymiar macierzy (6x6) i przechodząc dalej do trafności z odstępstwem o 1 na poziomie ≈ 96% widzimy, że taki wynik jest jak najbardziej satysfakcjonujący. **MAE** otrzymałem na poziomie **0.4708** co również jest zadowalające. Podsumowując, algorytm ten z małą tolerancją błędu działa na dość zadowalającym poziomie i spisał się dobrze.

Szacowanie – MLP

Problem szacowania postanowiłem rozwiązać przy użyciu wielowarstwowych perceptronów. Pracę nad nim rozpocząłem od standaryzacji Min-Max. Następnie na odpowiednio spreparowanych danych zbudowałem sieć neuronową składającą się z 8 warstw ukrytych, ze stałą uczenia alpha na poziomie 0.0001, funkcją aktywacji jako tangens hiperboliczny i z ziarnem generatora liczb pseudolosowych ustawionym na mój numer indeksu, nauczyłem ją na zestandaryzowanych danych uczących i przeprowadziłem predykcję. Kolejnym krokiem było przeprowadzenie destandaryzacji wyników predykcji oraz odpowiednie przygotowanie ich, aby można było stworzyć poprawną macierz pomyłek. Dzięki niej ponownie mogłem policzyć trafność oraz trafność z odstępstwem o 1. Otrzymałem następujące wyniki:

```
[[  0   0   3   0   0   0]
 [  1   0  10   5   0   0]
 [  0   4 141  56   3   0]
 [  0   2  55 124  11   0]
 [  0   0   1  39  20   0]
 [  0   0   0   1   4   0]]

Trafność : 285/480 = 0.59375
```

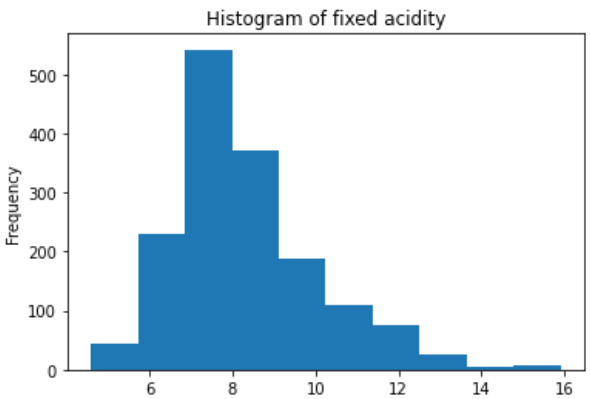
```
[[  0   0   3   0   0   0]
 [  1   0  10   5   0   0]
 [  0   4 141  56   3   0]
 [  0   2  55 124  11   0]
 [  0   0   1  39  20   0]
 [  0   0   0   1   4   0]]

Trafność z odstepstwem o 1: 465/480 = 0.96875
```

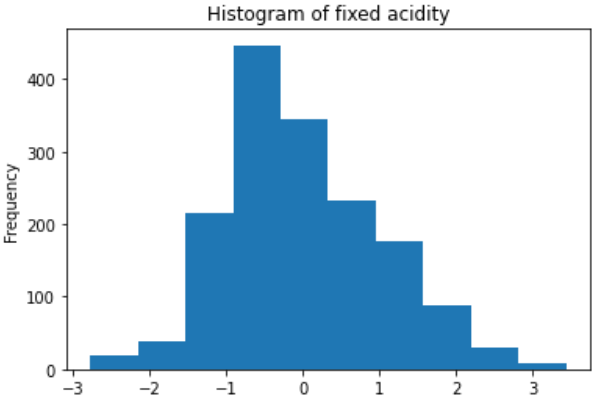
Ponownie trafność na poziomie ≈ 59% nie jest zadowalająca jednak patrząc na duży wymiar macierzy (6x6) i przechodząc dalej do trafności z odstępstwem o 1 na poziomie ≈ 97% widzimy, że taki wynik jest jak najbardziej satysfakcjonujący. **MAE** otrzymałem na poziomie **0.4375** co również jest zadowalające. Podsumowując, podobnie do KNN algorytm ten z małą tolerancją błędu również działa na dość zadowalającym poziomie i spisał się dobrze.

Otrzymane wyniki pokazują, że algorytm ten osiągnął znacznie lepszą trafność i MAE, co za tym idzie lepiej przewiduje niż algorytm KNN. Jeśli miałbym dokonać wyboru, użycie MLP byłoby lepszym rozwiązaniem.

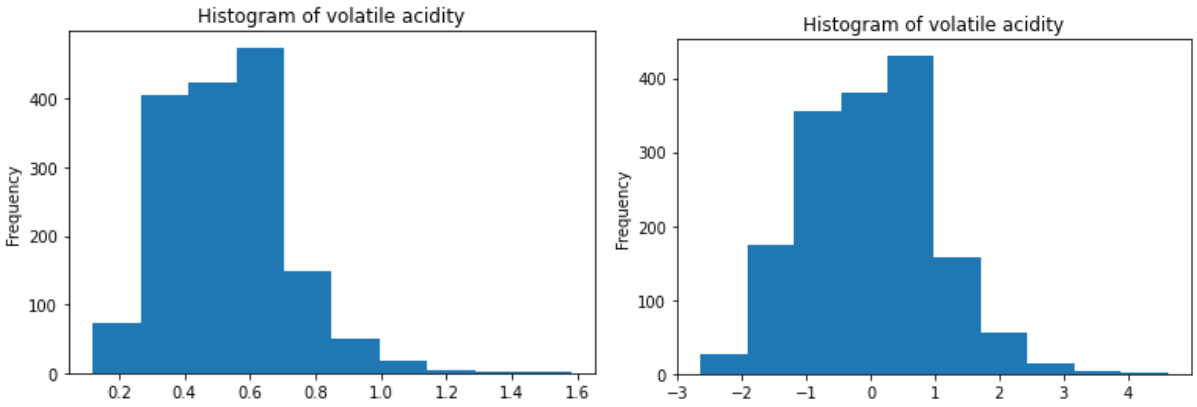
Grupowanie – Algorytm centroidów

Problem grupowania postanowiłem rozwiązać przy pomocy algorytmu centroidów /K-średnich. Prace nad nim rozpocząłem od stworzenia histogramów dla zmiennych zarówno przed przekształceniami (standaryzacją zmiennych (ln(x)+1)), jak i po.

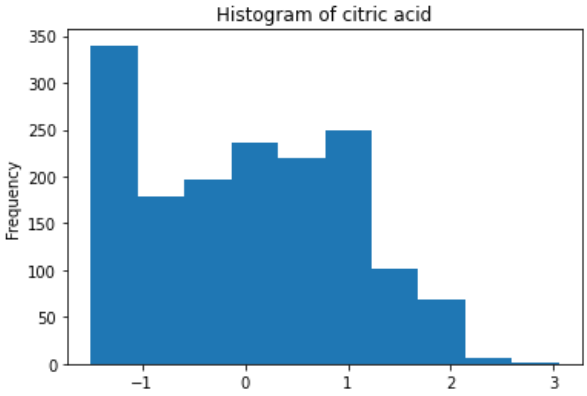


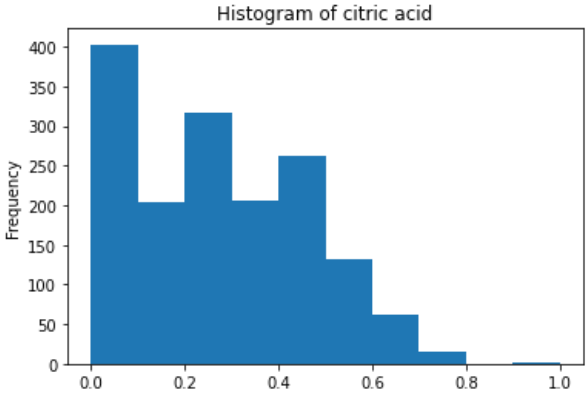


Przed przekształceniami histogram ten był prawostronnie skośny, niesymetryczny, ale delikatnie przypominający rozkład normalny. Po przekształceniach wciąż prawostronnie skośny, jednak bardziej wygładzony niż poprzedni.

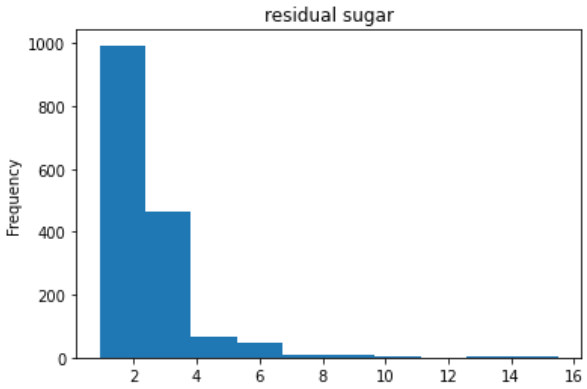


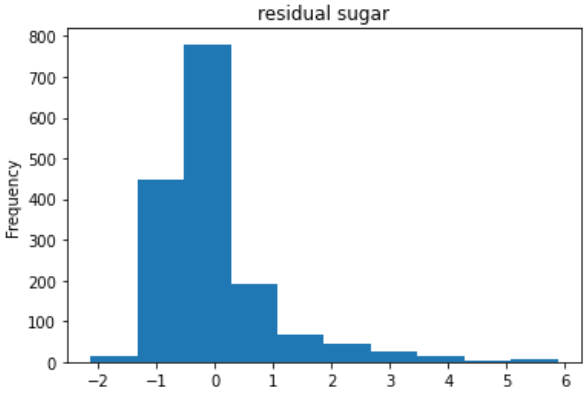
W tym wypadku widzimy histogram niesymetryczny. Po przekształceniu możemy zauważyć wyraźną poprawę, jednak wciąż nie jest to w pełni zadowalające.



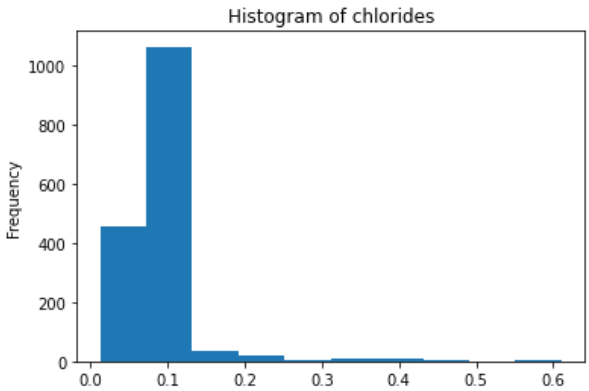


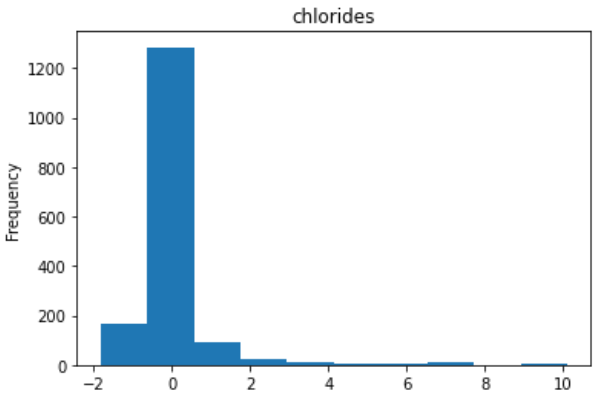
Kolejny histogram jest bardzo niesymetryczny, po przekształceniach delikatnie się wygładza, ale wciąż nie w żaden sposób nie przypomina układu normalnego.



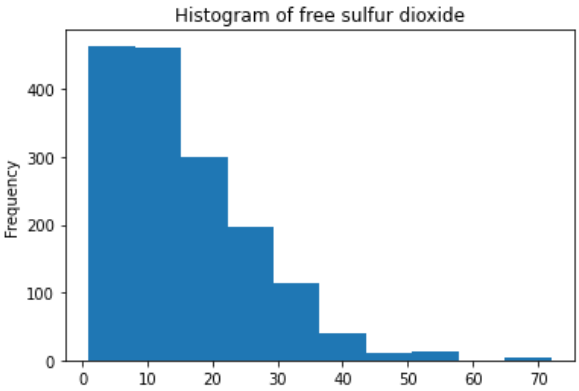


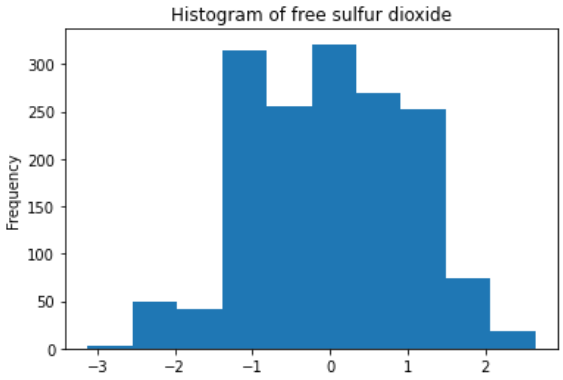
W tym wypadku mamy do czynienia z histogramem niesymetrycznym, prawostronnie skośnym, jednak po przekształceniu widzimy znaczną poprawę, mimo że skośność pozostała.



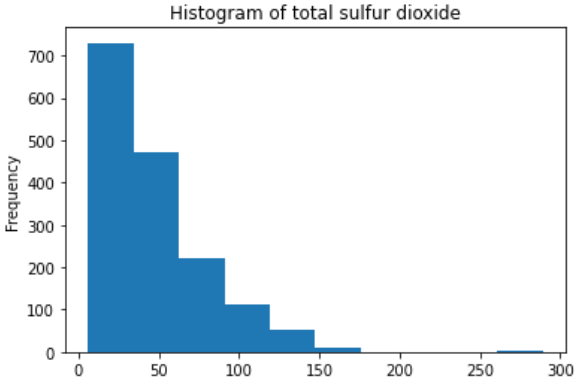


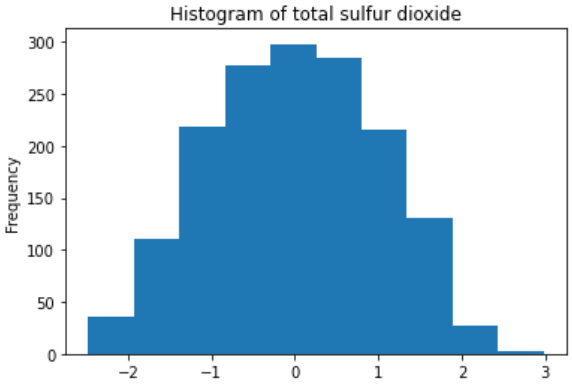
Tutaj widzimy histogram niesymetryczny, prawostronnie skośny, który po przekształceniu bardzo ładnie wygładza się mimo niewielkiej liczbie obserwacji skrajnych powodujących tą skośność.



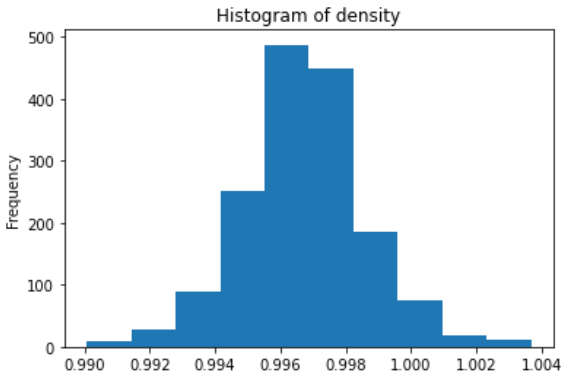


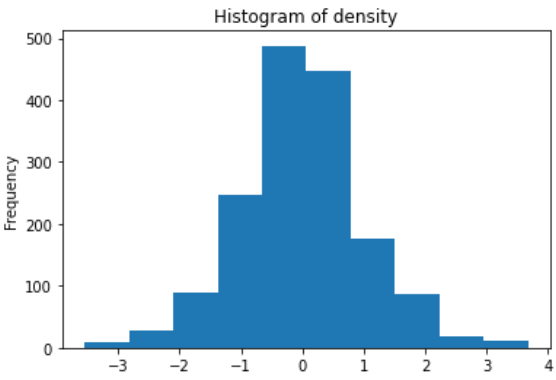
Histogram ten jest całkowicie niesymetryczny, prawostronnie skośny, jednak po przekształceniu możemy zaobserwować wyraźną zmianę. Daleko jest mu do dobrego histogramu, jednak symetria zaczęła się pojawiać i zniwelowaliśmy skośność.



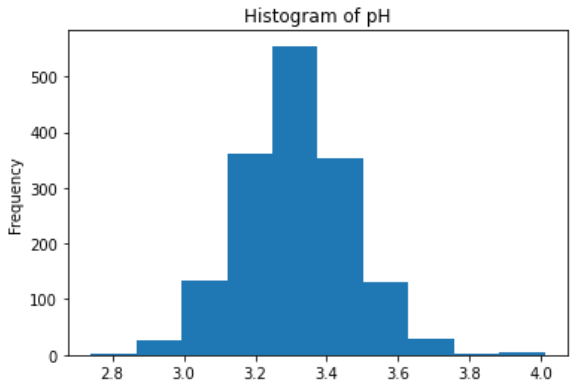


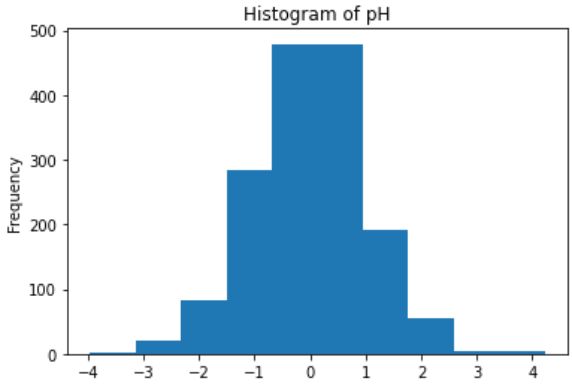
Sytuacja bardzo podobna do poprzedniego histogramu, jednak z lepszym zakończeniem. Pierwotnie całkowicie niesymetryczny i prawostronnie skośny, natomiast po przekształceniu otrzymujemy niemalże idealny histogram. Pojawiła się piękna symetryczność i skośność została całkowicie zniwelowana.



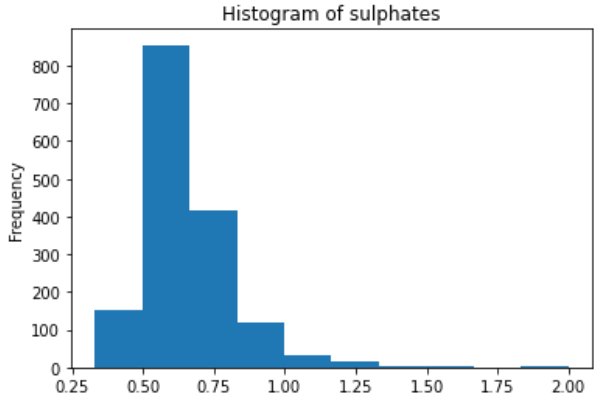


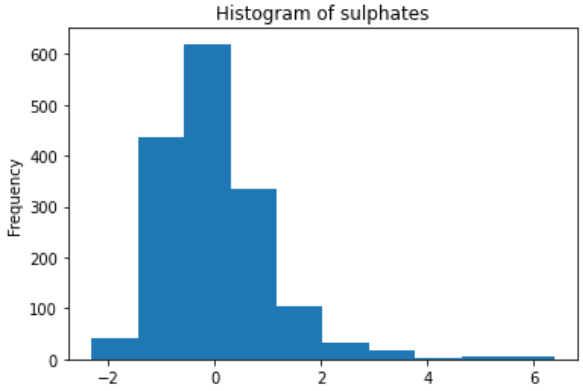
W tym wypadku mamy do czynienia z niezauważalną zmianą przed i po przekształceniu. Histogramy zbliżone do rozkładu normalnego, lekko niesymetryczne, ale zadowalające.



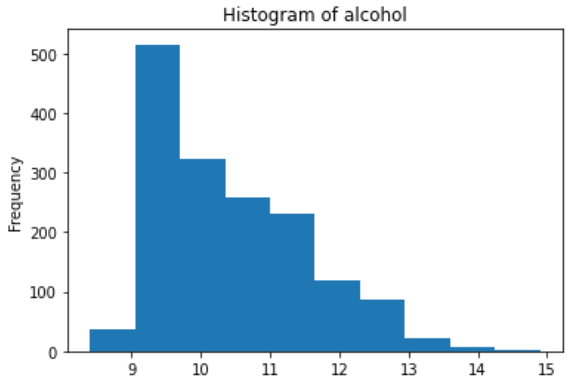


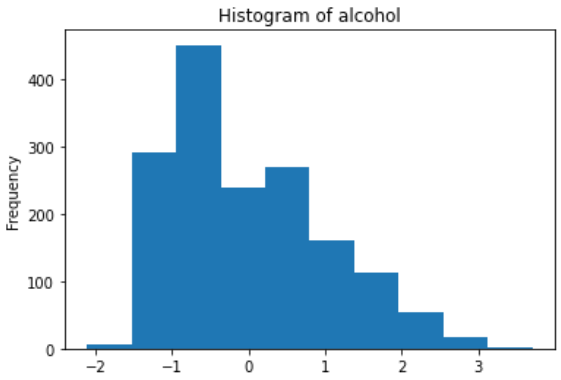
Przed przekształceniem histogram niemalże idealny, po przekształceniu ulegl delikatnemu pogorszeniu, jednak wciąż bardziej zadowalający niż wiele innych.





Na samym początku mamy do czynienia z histogramem niesymetrycznym, prawostronnie skośnym, jednak po przekształceniu sytuacja się poprawia. Pojawia się delikatna symetria, natomiast skośność wciąż pozostaje.





Na ostatnim histogramie widzimy duża niesymetryczność i prawostronną skośność. Po przekształceniach histogram delikatnie się poprawia, ale wciąż nie jest to zadowalający wynik.

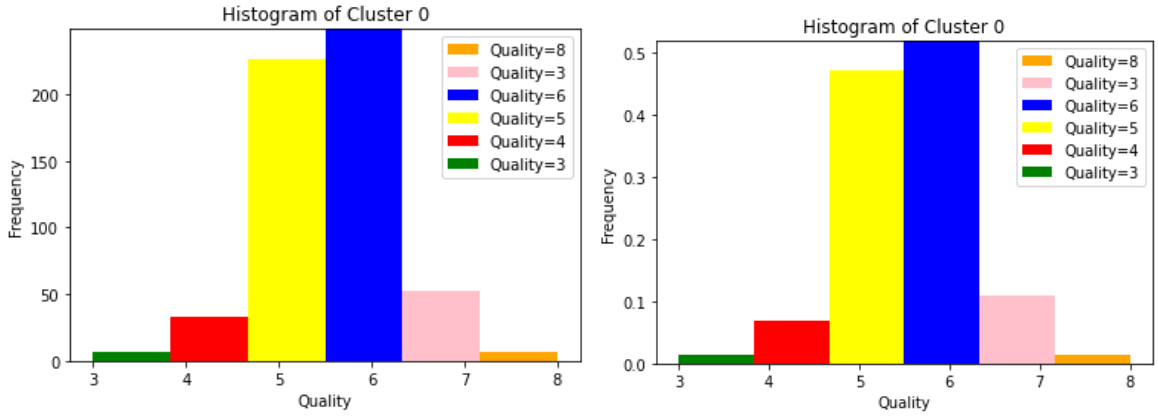
Po wykonaniu wszystkich histogramów przeszedłem do algorytmu K-średnich. Postanowiłem podzielić dane na 3 klastry (przy 4 jeden z nich był bardzo nieliczny (zawierał jedynie 29 obserwacji)) ustanwiając ziarno generatora liczb pseudolosowych na mój numer indeksu. Następnie po otrzymaniu wyników przeszedłem do ich analizy.

Opis pierwszego klastra

Pierwszy klaster okazał się dość liczny. Zawierał 576 obserwacji. Zwrócił mi większość składników poniżej średniej poza alkoholem, pH i lotną kwasowością. Po zaczerpnięciu informacji z internetu, stwierdziłem, że będą to wina wytrawne ze względu na najmniejszą zawartość siarczanów, tlenków siarki i cukrów.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000	576.000000
mean	-0.764147	0.543017	-0.878656	-0.290831	-0.215961	-0.269841	-0.454336	-0.579722	0.701650	-0.319215	0.193944
std	0.614912	0.904520	0.619745	0.696647	0.550839	0.881336	0.771260	0.843097	0.825302	0.729159	1.039419
min	-2.769397	-2.339214	-1.497274	-1.592065	-1.272557	-2.008867	-2.279303	-3.475170	-1.523412	-2.304047	-1.425495
25%	-1.054196	-0.044068	-1.431973	-0.698491	-0.490632	-0.897333	-1.011650	-1.074275	0.204696	-0.739474	-0.677107
50%	-0.696331	0.514070	-1.053247	-0.442852	-0.251089	-0.247127	-0.391953	-0.527930	0.588773	-0.398482	0.023839
75%	-0.359909	1.013077	-0.466898	-0.089701	-0.061052	0.311420	0.042224	0.012833	1.217410	0.061312	0.864336
max	0.761234	4.618472	1.828842	3.544551	3.940962	2.073160	1.521305	1.405880	4.220567	2.468416	3.059447

Następnie stworzyłem histogramy (zwykly i znormalizowany) dla wskazanej grupy wybierając zmienna quality.



Możemy zauważyć, że grupie win wytrawnych zdecydowaną przewagę mają wina średniej jakości, jednak spotkamy się tutaj też z winami rewelacyjnymi (ocena 8) jak i “tanimi jabolami” ocena (3).

Opis drugiego klastra

Klaster ten zawierał najmniej obserwacji (różnica 100 miedzy najwiekszym) jednak wciąż był bardzo liczny – 475 obserwacji. Składniki zawarte w nim wskazały:

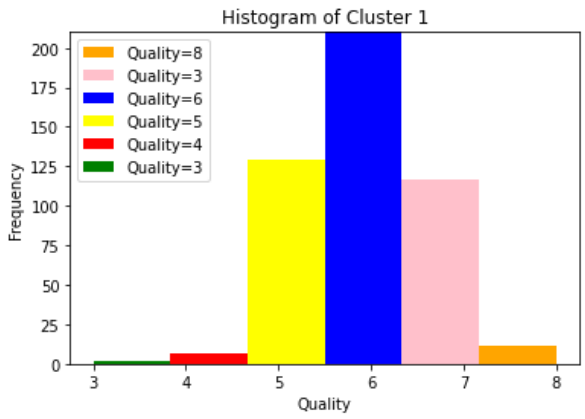
- Stała kwasowość powyżej średniej
- Lotną kwasowość poniżej średniej
- Kwas cytrynowy powyżej średniej
- Cukier resztkowy poniżej średniej
- Chlorki poniżej średniej

- Wolny dwutlenek siarki poniżej średniej
- Całkowity dwutlenek siarki poniżej średniej
- Gęstość delikatnie powyżej średniej
- PH zdecydowanie poniżej średniej
- Siarczany delikatnie powyżej średniej
- Alkohol delikatnie powyżej średniej

Na tej podstawie nazwałem tą grupą winami półsłodkimi, ze względu na wciąż wysoką zawartość siarczanu i ponad normalną zawartość cukru.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	476.000000	476.000000	476.000000	476.000000	476.000000	476.000000	476.000000	476.000000	476.000000	476.000000	476.000000
mean	1.025600	-0.694755	1.011611	0.075838	0.300982	-0.595404	-0.558698	0.451202	-0.807030	0.557607	0.285145
std	0.813643	0.768397	0.572000	0.980189	1.585854	0.847812	0.803917	0.997740	0.790849	1.186172	1.015434
min	-1.203981	-2.644648	-0.812386	-1.427358	-1.173503	-3.120402	-2.474994	-1.774859	-3.957395	-1.605257	-2.109913
25%	0.430752	-1.280750	0.663863	-0.568467	-0.394546	-1.111464	-1.090885	-0.236308	-1.250501	-0.265079	-0.472264
50%	1.022518	-0.764808	1.075599	-0.203719	-0.132151	-0.708455	-0.609463	0.359942	-0.779221	0.378016	0.216213
75%	1.604062	-0.218735	1.336405	0.334222	0.274002	0.000070	0.042224	1.088236	-0.250162	1.043350	1.042747
max	3.434380	1.909687	3.051676	5.873725	10.103357	1.799545	1.883506	3.415415	1.464962	6.377026	3.703973

Kolejno wykonałem znów histogramy (zwykły i znormalizowany) dla tej samej zmiennej quality, tylko następnego klastra.



Możemy zauważyć, że w grupie win półsłodkich ponownie mamy większość win średnich jakościowo, ale należy zwrócić uwagę, że zdecydowana większość win bardzo dobrych i rewelacyjnych zalicza się do niego. Spotkamy się również ze słabymi winami, ale w bardzo małej ilości.

Opis trzeciego klastra

Trzeci klaster zawierał 547 obserwacji. Zwrócił następujące dane:

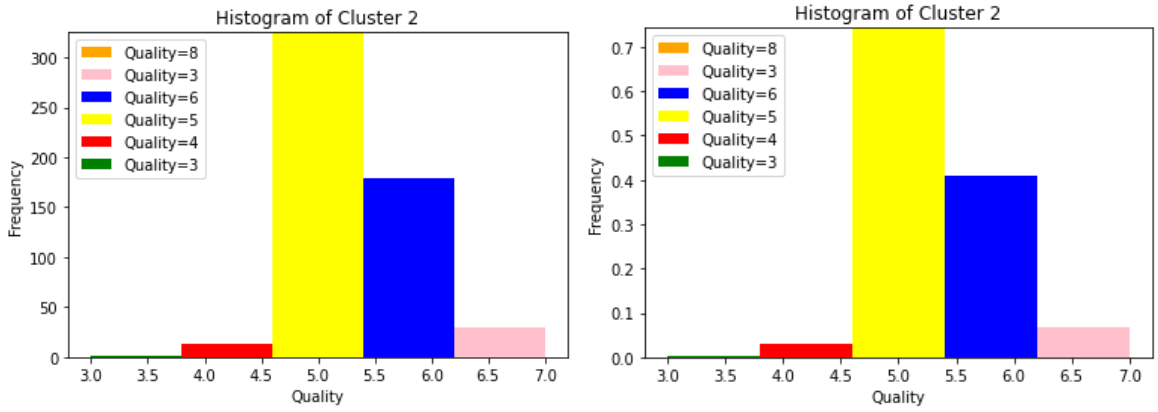
- Stała kwasowość poniżej średniej
- Lotną kwasowość delikatnie powyżej średniej
- Kwas cytrynowy delikatnie powyżej średniej
- Cukier resztkowy powyżej średniej
- Chlorki delikatnie poniżej średniej
- Wolny dwutlenek siarki powyżej średniej
- Całkowity dwutlenek siarki powyżej średniej

- Gęstość delikatnie powyżej średniej
- PH zdecydowanie delikatnie poniżej średniej
- Siarczany delikatnie poniżej średniej
- Alkohol delikatnie poniżej średniej

Grupę tą nazwałem winami słodkimi ze względu na większą zawartość tlenków siarki i delikatnie powiększoną zawartość cukru.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	547.000000	547.000000	547.000000	547.000000	547.000000	547.000000	547.000000	547.000000	547.000000	547.000000	547.000000
mean	-0.087819	0.032770	0.044934	0.240256	-0.034504	0.802269	0.964602	0.217820	-0.036570	-0.149092	-0.452360
std	0.643203	0.905735	0.728569	1.195646	0.540150	0.674212	0.563775	0.856977	0.760697	0.862432	0.760397
min	-1.846576	-2.339214	-1.497274	-2.135272	-1.824297	-1.358661	-0.443380	-3.544303	-2.857391	-1.833141	-1.992862
25%	-0.559335	-0.640463	-0.383353	-0.442852	-0.298819	0.311420	0.519131	-0.281371	-0.579650	-0.739474	-0.991666
50%	-0.167321	0.127163	0.019454	-0.089701	-0.108429	0.864408	0.940729	0.187727	-0.054317	-0.331572	-0.677107
75%	0.373730	0.621528	0.520670	0.433035	0.115836	1.274824	1.369920	0.664563	0.461332	0.315422	-0.073618
max	1.873710	3.388483	2.286894	5.851201	4.001636	2.648268	2.982484	3.674262	1.832229	4.931614	2.217009

Tutaj również wykonałem dwa histogramy (zwykły i znormalizowany) dla zmiennej quality i ostatniego klastra.



Widzimy tutaj, że wśród win słodkich, nie mamy ani jednego wina rewelacyjnego (ocena 8), ale również wina z najniższymi ocenami (3 i 4) mają tutaj niewielki wkład. Zdecydowana większość to wina z oceną 5, jednak znajdziemy tutaj tez wina lepsze pod względem oceny.