

# Multidimensional Data Visualization

## 1. Zadanie 1

### 4.1. Wybór zbioru danych

Do analizy wybrałem zbiór danych "Wine quality - red" z bazy UCI Machine Learning Repository dostępny wśród zbiorów danych dostarczonych wraz z narzędziem do analizy danych Orange.

Zbiór ten zawiera informacje o chemicznych cechach win czerwonych oraz ich ocenach jakości.

W zbiorze zawartych jest 11 cech numerycznych, opisujących różne właściwości chemiczne wina. Są to:

- **fixed acidity** (stała kwasowość),
- **volatile acidity** (lotna kwasowość),
- **citric acid** (kwas cytrynowy),
- **residual sugar** (cukry resztkowe),
- **chlorides** (chlorki),
- **free sulfur dioxide** (wolny dwutlenek siarki),
- **total sulfur dioxide** (całkowity dwutlenek siarki),
- **density** (gęstość),
- **pH** (wartość pH),
- **sulphates** (siarczany),
- **alcohol** (zawartość alkoholu).

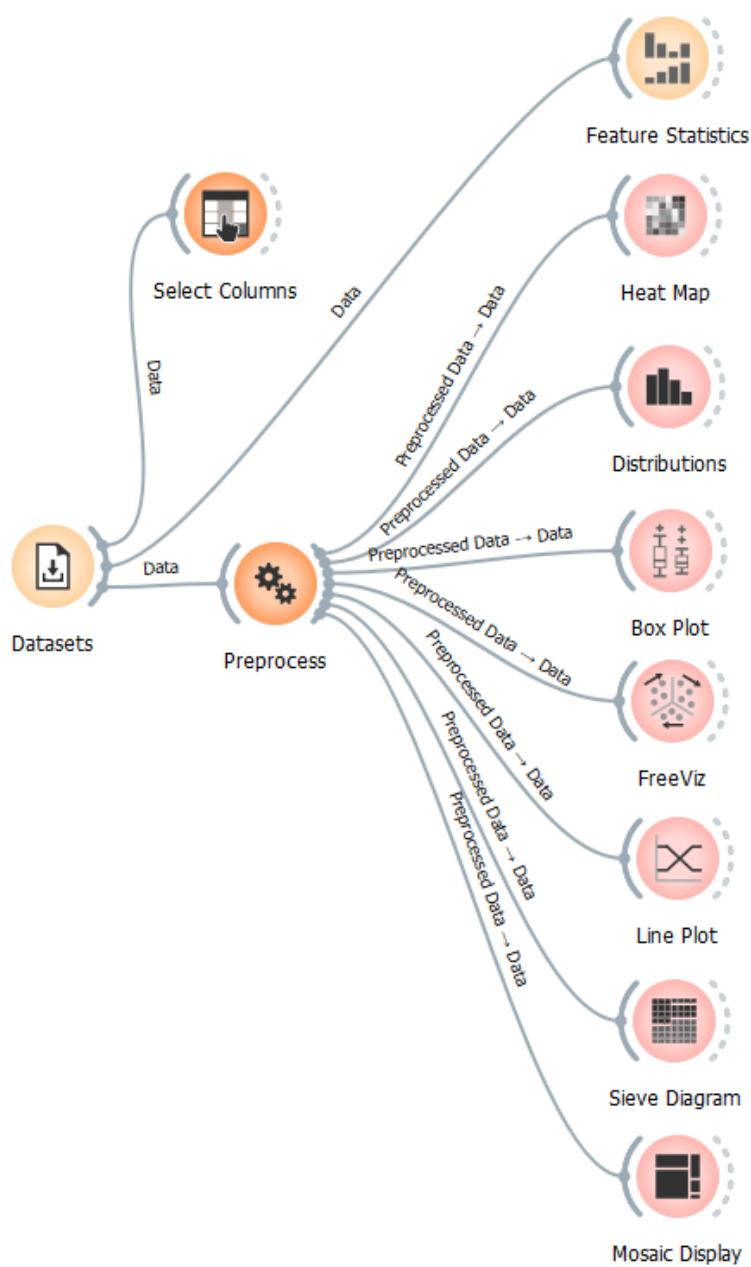
Dodatkowo, zawiera kolumnę z oceną jakości wina na skali od 0 do 10. Zbiór danych jest złożony i dobrze udokumentowany, dlatego nadaje się do analizy w tym zadaniu.

### 4.2. Obserwacje na podstawie widgetów wizualizacji

Do obserwacji cech wykorzystałem widoczne na zamieszczonym na następnej stronie diagramie widgety cech. Są to:

- **Feature Statistics:** Pozwala na sprawdzenie podstawowych statystyk każdej cechy, takich jak średnia, odchylenie standardowe, minima i maksima.
- **Heat Map:** Umożliwia analizę korelacji między cechami i identyfikację tych, które mają wysokie współczynniki korelacji z jakością wina.
- **Distributions:** Wizualizuje rozkład wartości każdej cechy, co pozwala na ocenę ich różnic między klasami jakości.

- **Box Plot:** Pokazuje rozkład i zakres wartości każdej cechy, co pomaga w identyfikacji cech, które różnicują klasy jakości.
- **FreeViz:** Umożliwia wizualizację wielowymiarowych danych i identyfikację cech, które najlepiej różnicują klasy jakości wina.
- **Line Plot:** Pozwala na analizę trendów i zmian w danych, co pomaga w identyfikacji istotnych cech.
- **Sieve Diagram:** Analizuje asocjacje między cechami a jakością wina, co pozwala na identyfikację cech o silnych powiązaniach.
- **Mosaic Display:** Wizualizuje wielowymiarowe zależności między cechami, co pomaga w identyfikacji cech istotnych dla jakości wina.



### 1.5.1. Feature Statistics

W wynikach dla zbioru danych "Wine quality - red" widzimy, że niektóre cechy mają szeroki zakres wartości i duże rozproszenie, co może sugerować ich istotność w analizie jakości wina.

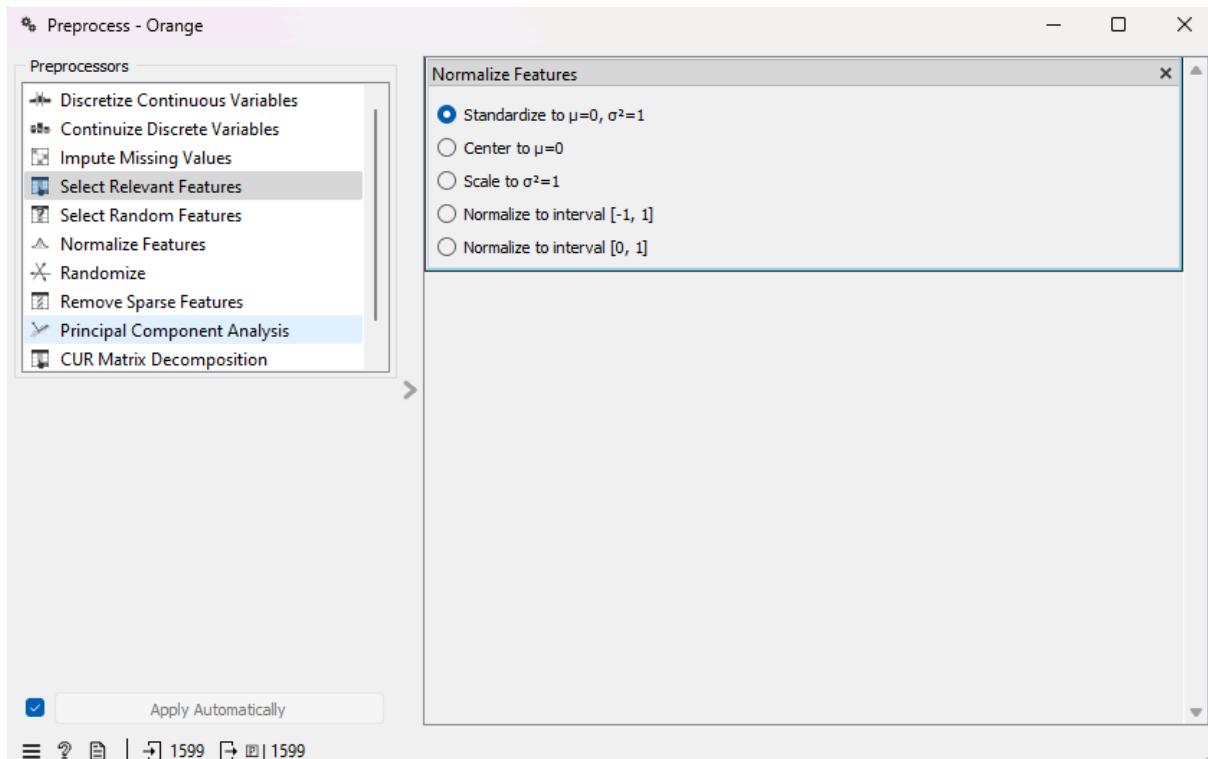
Na przykład:

- **Fixed acidity:** Ma średnią wartość 8.32 i rozproszenie 0.209, co wskazuje na różnorodność w stałej kwasowości próbek wina.
- **Volatile acidity:** Średnia 0.52782 i rozproszenie 0.33914 wskazują na zmienność w lotnej kwasowości, co może wpływać na jakość wina, ponieważ wysokie poziomy lotnej kwasowości mogą obniżać jakość.
- **Alcohol:** Ma średnią wartość 10.423 i najniższe rozproszenie 0.10221, co sugeruje, że zawartość alkoholu jest stosunkowo stała, ale jej wpływ na jakość wina może być znaczący.
- **Citric acid:** Ma niską średnią 0.2710, ale najwyższe rozproszenie 0.7187, co sugeruje, że jego obecność może znacząco wpływać na różnice w jakości wina.

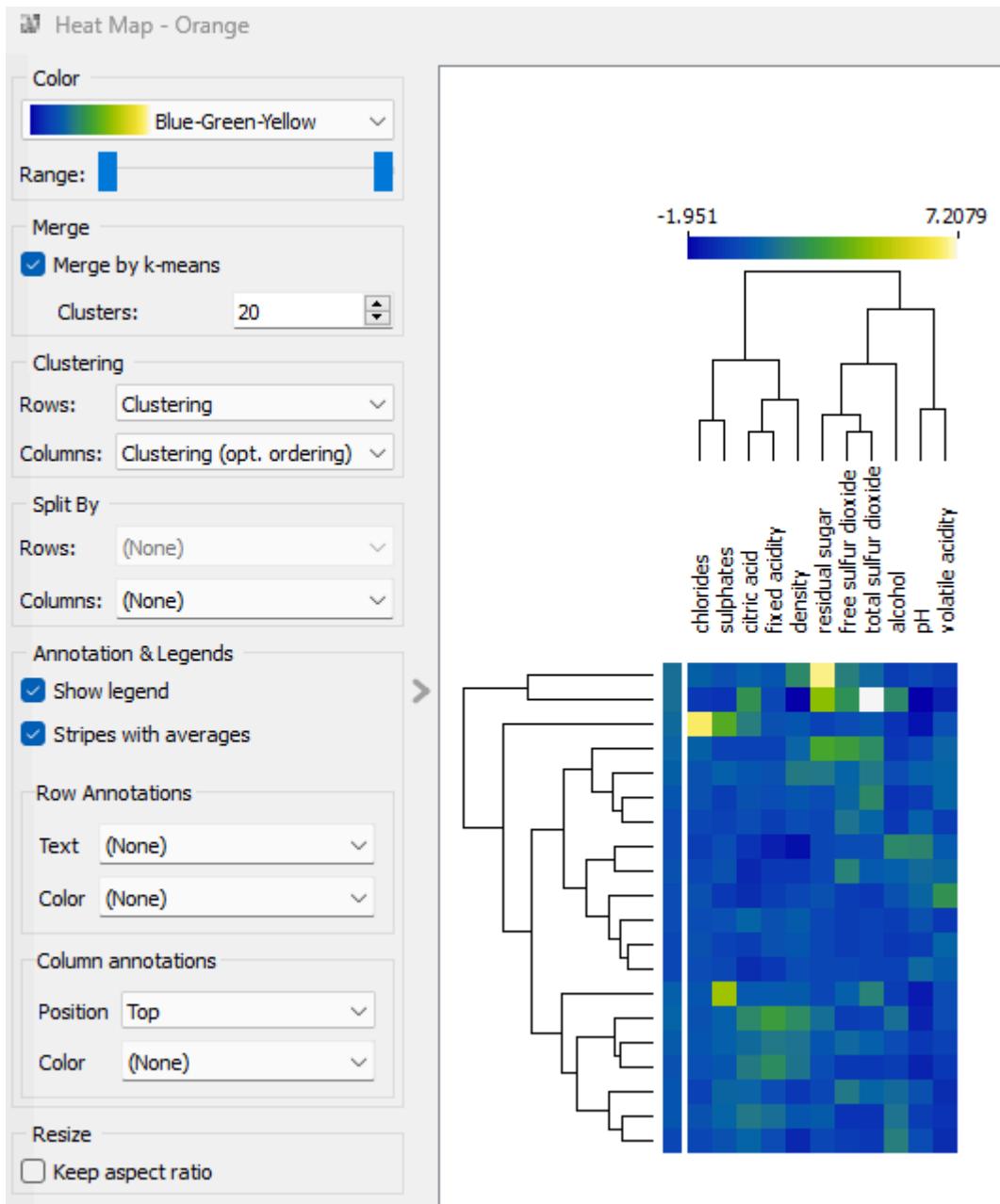
Na podstawie wyników widgetu **Feature Statistics**, cechy takie jak **volatile acidity**, **fixed acidity**, **alcohol**, i **citric acid** mogą być najbardziej istotne. Wysokie rozproszenie wskazuje na dużą zmienność tych cech w danych, co sugeruje ich potencjalny wpływ na jakość wina.



Na podstawie statystyk cech, możemy zauważyc, że cechy charakteryzują się dużym rozrzutem, dlatego wykorzystamy węzeł preprocessingu, w celu normalizacji danych. Wybrałem standaryzację, która zmniejszy rozproszenie poszczególnych cech.



### 1.5.2. Heat Map



Na podstawie tej mapy ciepła można wnioskować, że najbardziej istotne cechy to:

- **Alcohol:** Wysokie wartości alkoholu są często skorelowane z wyższą jakością wina.
- **Volatile acidity:** Wysokie wartości lotnej kwasowości mogą obniżać jakość wina.
- **Citric acid:** Kwas cytrynowy, który tworzy grupę z innymi ważnymi cechami, również może mieć znaczący wpływ na jakość wina.
- **Sulphates i chlorides:** Ich grupowanie i zmienność w wartościach sugerują wpływ na jakość.

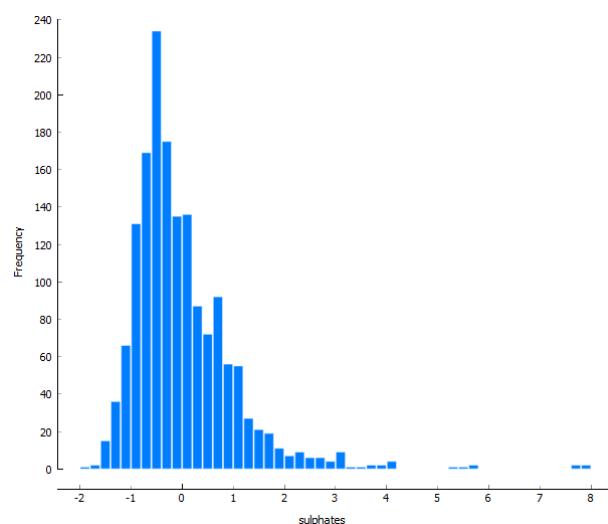
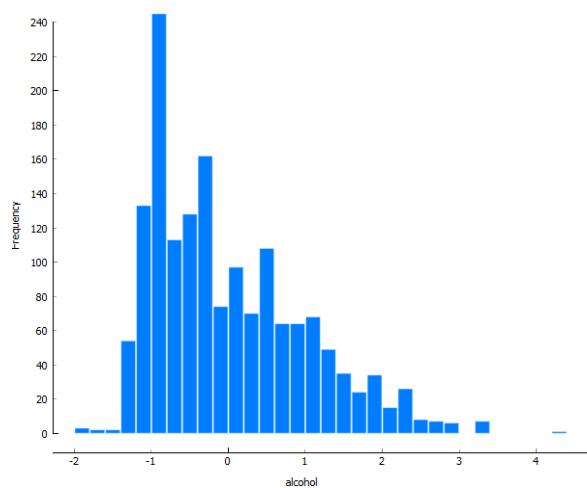
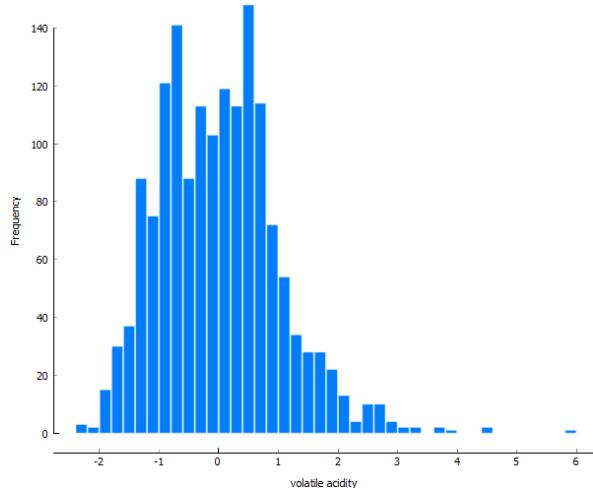
### 1.5.3. Distributions

Widget **Distributions** pozwala na dokładniejszą analizę rozkładu wartości poszczególnych cech. Dzięki niemu można zaobserwować, jak wartości cech są rozproszone, oraz czy są symetryczne, czy też wykazują skłonności do określonych wartości.

Przykłady Istotnych Cech:

- **Alcohol:** Można zaobserwować, że rozkład wartości alkoholu jest skoncentrowany wokół pewnej wartości, z wyraźnym pikiem, co sugeruje, że większość próbek wina ma podobną zawartość alkoholu.
- **Volatile acidity:** Rozkład lotnej kwasowości pokazuje, że większość próbek ma niższe wartości, ale są też obecne wartości ekstremalne, co może negatywnie wpływać na jakość wina.
- **Sulphates:** Rozkład siarczanów wykazuje dużą zmienność, co sugeruje, że ich wpływ na jakość może być znaczący.

Dzięki widgetowi **Distributions** możemy dokładniej zbadać, jak wartości poszczególnych cech są rozłożone w próbkach wina, co może pomóc w identyfikacji cech najbardziej wpływających na jakość wina. Cechy takie jak **alcohol**, **volatile acidity**, i **sulphates** wykazują interesujące wzorce rozkładu, które mogą być kluczowe dla dalszej analizy.

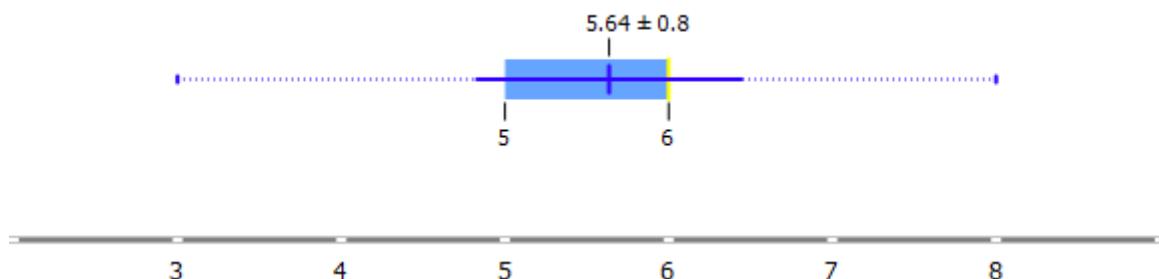


#### 1.5.4. Box Plot

Widget Box Plot pozwala na szczegółową analizę rozkładu cech oraz ich wpływu na jakość wina. Poniżej przedstawiono wyniki dla najciekawszych cech:

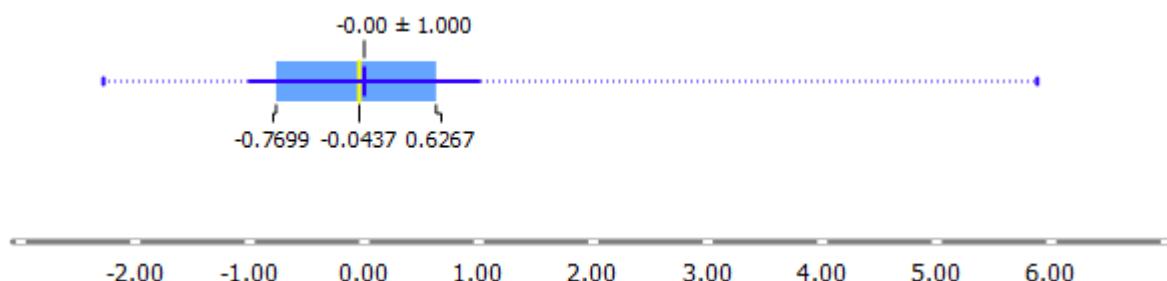
##### **Quality:**

Średnia wartość jakości wynosi 5.64, z zakresem od 3 do 8. Większość próbek ma wartość jakości między 5 a 6, co sugeruje, że większość win w zbiorze jest średniej jakości. Tylko niewielka liczba próbek osiąga wyższą jakość (7-8).



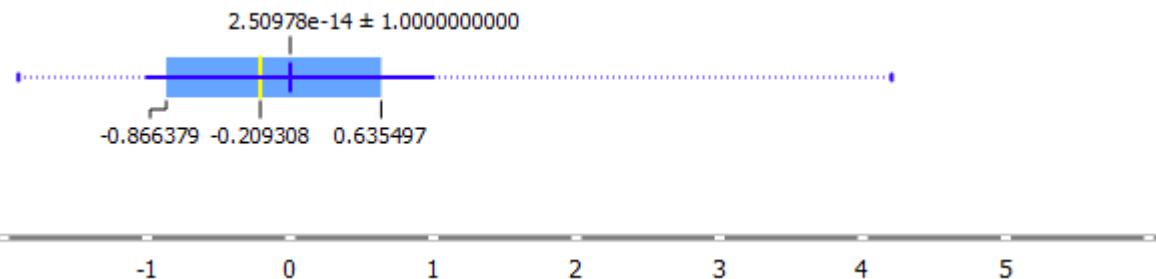
##### **Volatile Acidity:**

Lotna kwasowość ma średnią wartość około 0, z zakresem od -2 do 5. Większość wartości mieści się w przedziale -0.7 do 0.5. Wysokie wartości lotnej kwasowości mogą negatywnie wpływać na jakość wina, ponieważ wyższa kwasowość jest często niepożądana w smaku.



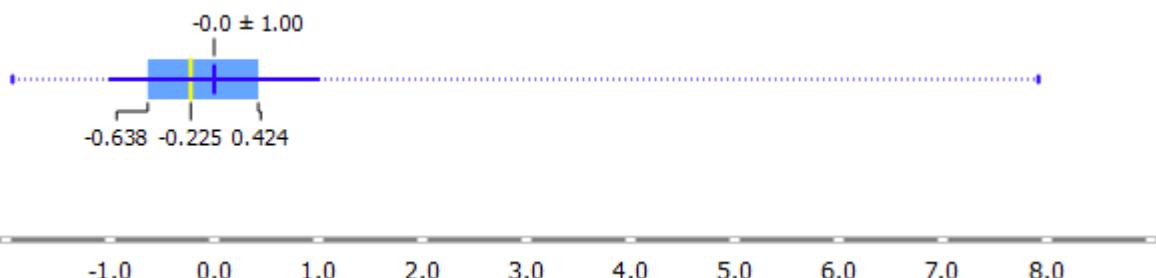
##### **Alcohol:**

Zawartość alkoholu wykazuje rozkład od -1 do 5, z większością próbek skupionych wokół wartości 0. Średnia wartość wynosi 2.50978e-14. Wyższa zawartość alkoholu jest często skorelowana z wyższą jakością wina, co sugeruje, że próbki z wyższymi wartościami alkoholu mogą być lepszej jakości.



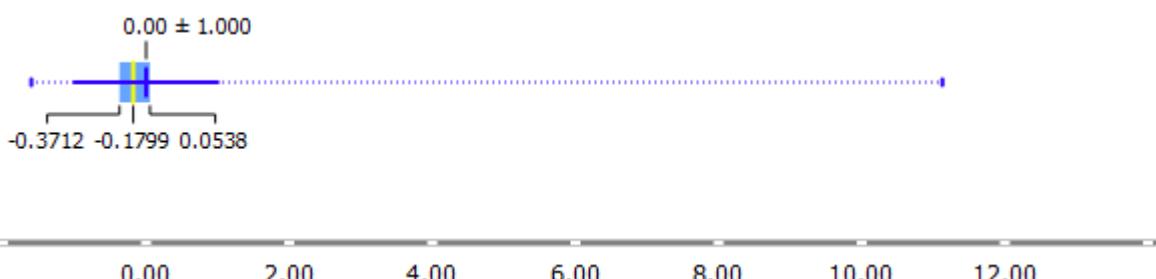
### Sulphates:

Rozkład siarczanów wykazuje znaczną zmienność, z większością próbek skupionych wokół wartości 0, ale z długim ogonem do wartości 8. Średnia wynosi 0, co sugeruje, że siarczany mają istotny wpływ na jakość wina, zwłaszcza w wyższych stężeniach. Wyższe stężenia siarczanów mogą przyczyniać się do lepszej jakości wina, ponieważ działają jako konserwanty.



### Chlorides:

Chlorki mają średnią wartość około 0, z zakresem od -0.5 do 10. Większość wartości znajduje się w przedziale -0.45 do 0.04. Wysoka zawartość chlorków może negatywnie wpływać na jakość wina, ponieważ nadmiar chlorków może prowadzić do nieprzyjemnego smaku.



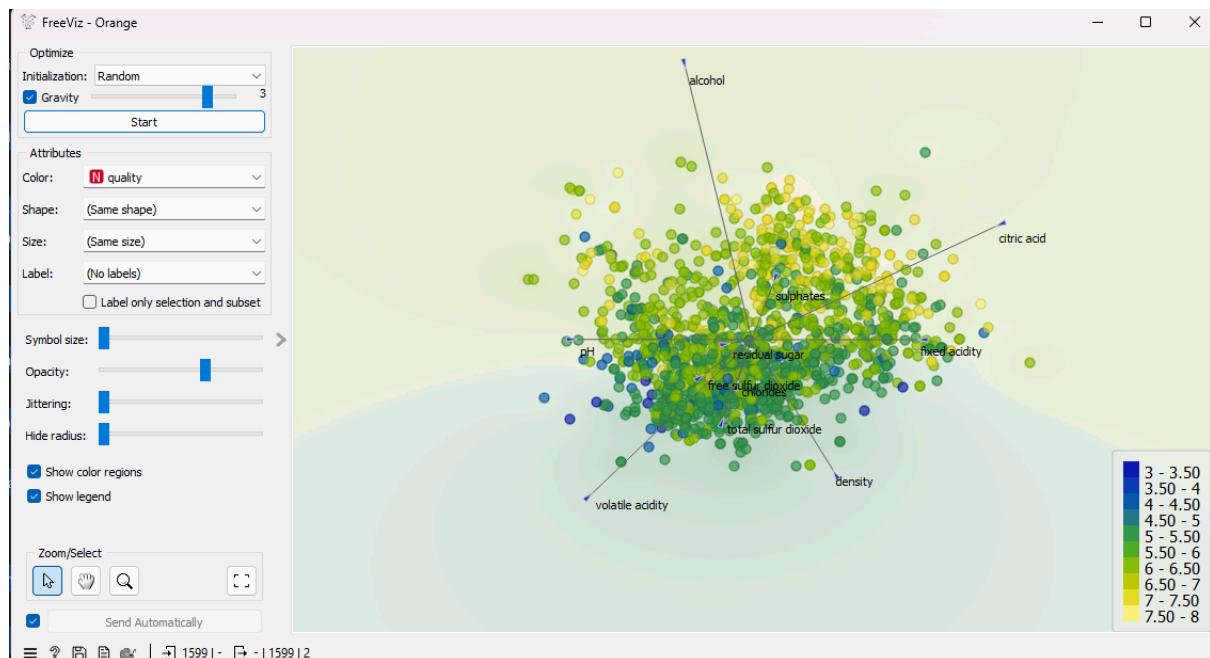
### 1.5.5. FreeViz

Na podstawie długości wektorów, cechy można uporządkować według ich wpływu na różnicowanie jakości wina:

1. **Alcohol:** Najdłuższy wektor, co sugeruje, że zawartość alkoholu ma największy wpływ na jakość wina. Wyższe wartości alkoholu są często skorelowane z wyższą jakością wina.
2. **Citric acid:** Drugi najdłuższy wektor, co wskazuje na istotny wpływ kwasu cytrynowego na jakość wina.
3. **Volatile acidity:** Kolejny długi wektor, wskazujący na znaczący wpływ lotnej kwasowości na jakość.
4. **Fixed acidity:** Długi wektor, sugerujący, że stała kwasowość również ma istotne znaczenie.

Kolejne wektory według długości, które już nie są aż tak długie, to:

5. **Density:** Wektor wskazujący na wpływ gęstości wina na jego jakość.
6. **pH:** Wektor również wskazujący na istotny wpływ.
7. **Total sulfur dioxide:** Sugeruje wpływ całkowitego dwutlenku siarki na jakość wina.
8. **Sulphates:** Mimo że wektor jest krótszy niż niektóre z powyższych, nadal wskazuje na znaczący wpływ.
9. **Pozostałe cechy:** Jak residual sugar, free sulfur dioxide i chlorides, mają krótsze wektory, co sugeruje mniejszy wpływ na różnicowanie jakości wina.



## 1.5.6. Line Plot

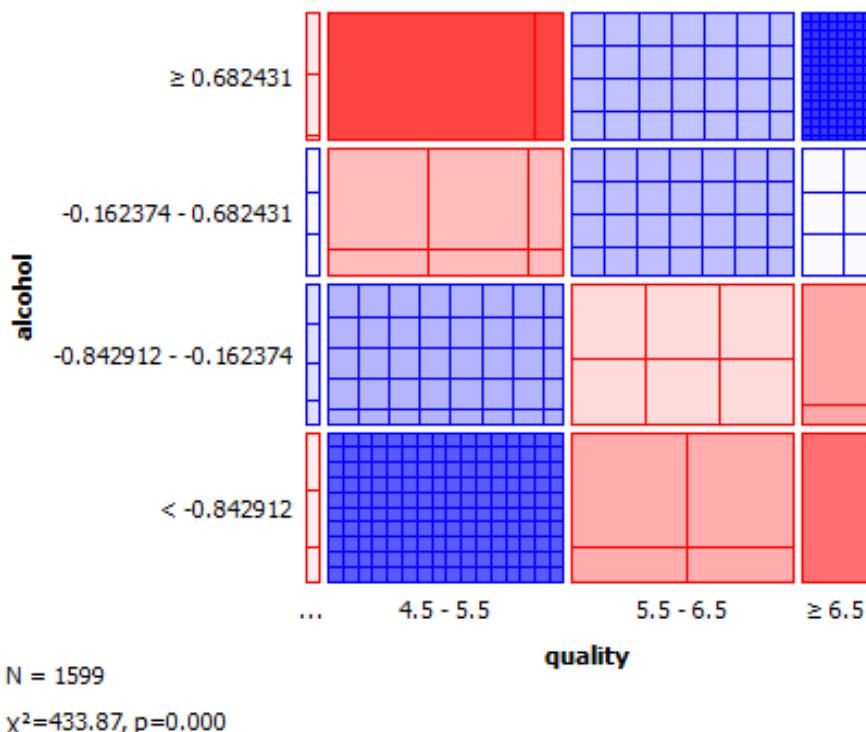
Możemy pominać, ponieważ nie daje istotnych w analizie wyników.

## 1.5.7. Sieve Diagram

Możemy pominać, ponieważ nie daje istotnych w analizie wyników.

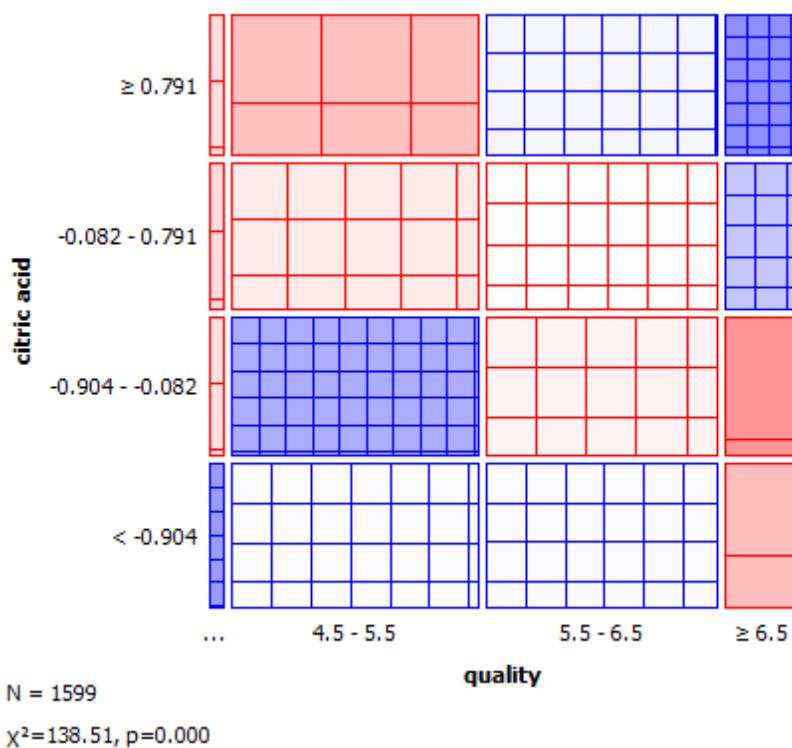
### Quality vs. Alcohol

- **Opis:** Diagram pokazuje, że wyższe wartości alkoholu ( $\geq 0.682431$ ) są silnie skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Czerwone prostokąty wskazują na nadreprezentację, a niebieskie na niedoszacowanie.
- **Wniosek:** Wyższa zawartość alkoholu jest istotnie skorelowana z wyższą jakością wina.



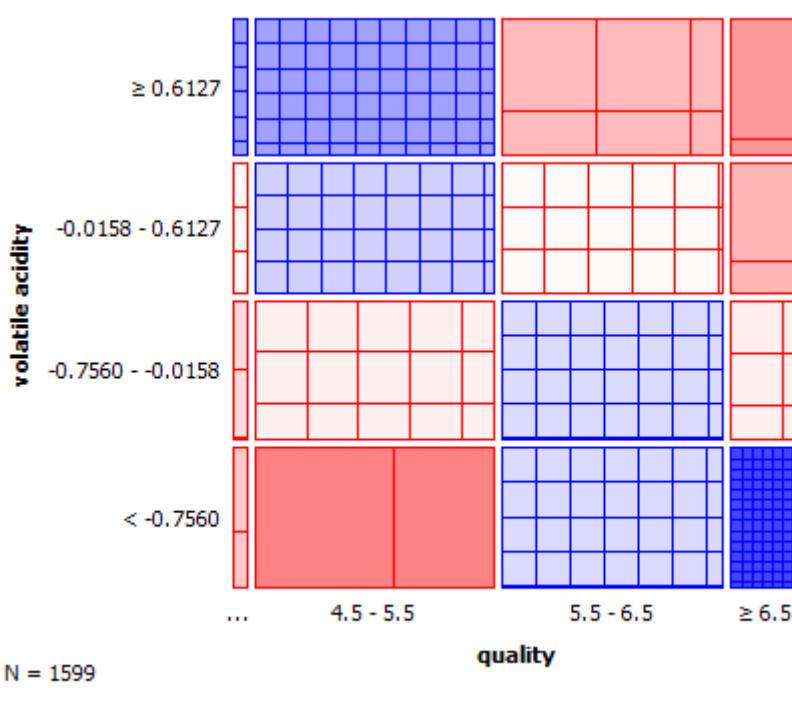
### Quality vs. Citric Acid

- **Opis:** Diagram pokazuje, że wyższe wartości kwasu cytrynowego ( $\geq 0.791$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Niskie wartości kwasu cytrynowego ( $< -0.904$ ) są związane z niższą jakością.
- **Wniosek:** Wyższa zawartość kwasu cytrynowego jest skorelowana z wyższą jakością wina.



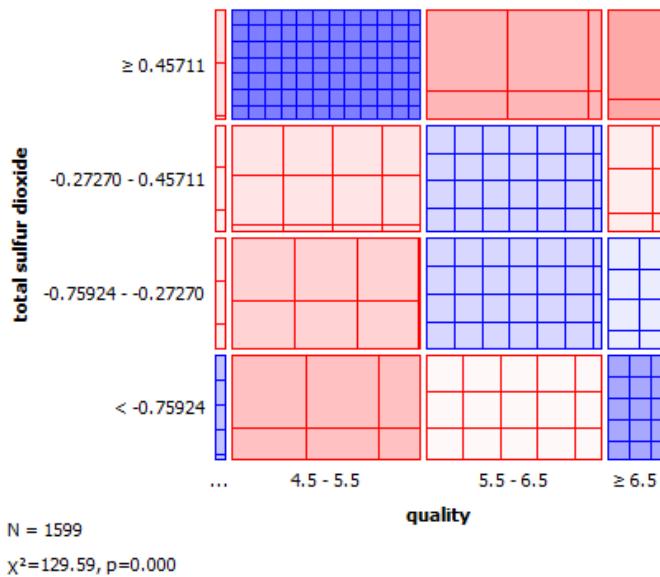
### Quality vs. Volatile Acidity

- **Opis:** Diagram wskazuje, że niższe wartości lotnej kwasowości ( $< -0.7560$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Wyższe wartości lotnej kwasowości są związane z niższą jakością.
- **Wniosek:** Niższa lotna kwasowość jest korzystna dla wyższej jakości wina.



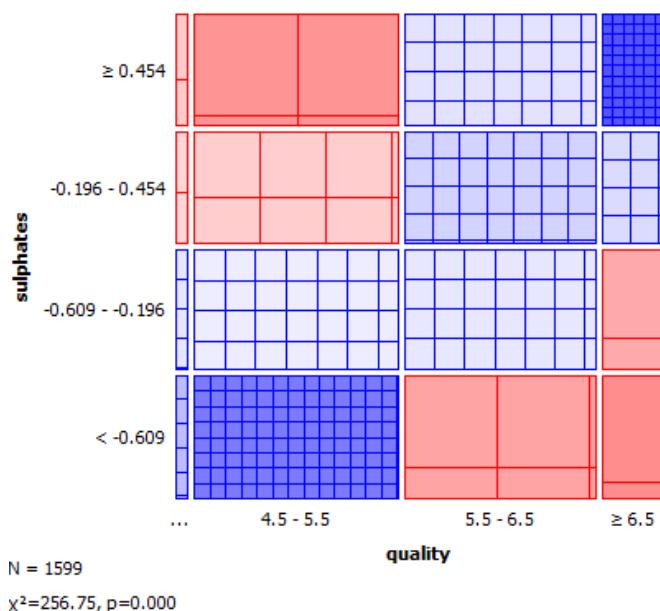
## Quality vs. Total Sulfur Dioxide

- **Opis:** Diagram pokazuje, że wyższe wartości całkowitego dwutlenku siarki ( $\geq 0.45711$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Niskie wartości są związane z niższą jakością.
- **Wniosek:** Wyższe wartości całkowitego dwutlenku siarki mogą być korzystne dla wyższej jakości wina.



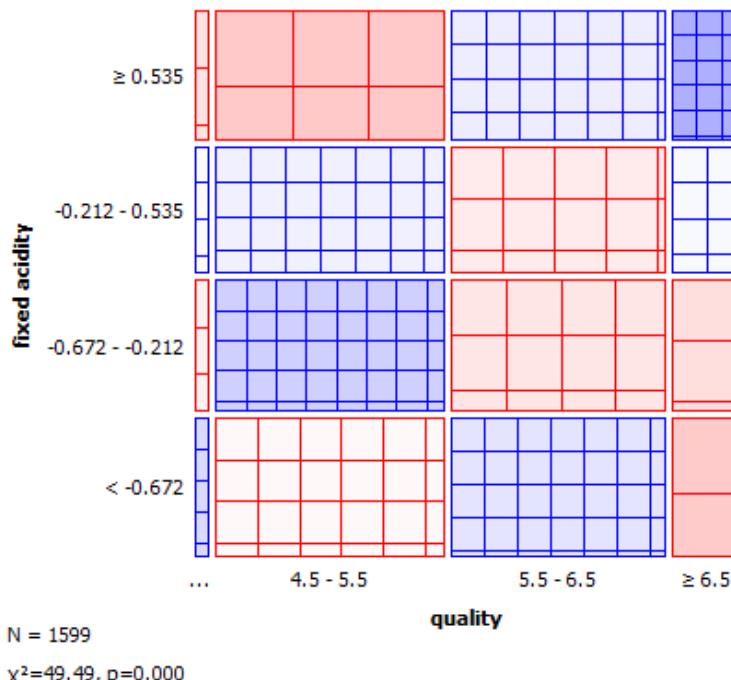
## Quality vs. Sulphates

- **Opis:** Diagram pokazuje, że wyższe wartości siarczanów ( $\geq 0.454$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Niższe wartości są związane z niższą jakością.
- **Wniosek:** Wyższe stężenie siarczanów jest korzystne dla wyższej jakości wina.



### Quality vs. Fixed Acidity

- **Opis:** Diagram pokazuje, że wyższe wartości stałej kwasowości ( $\geq 0.535$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Niższe wartości są związane z niższą jakością.
- **Wniosek:** Wyższa stała kwasowość może być skorelowana z wyższą jakością wina.



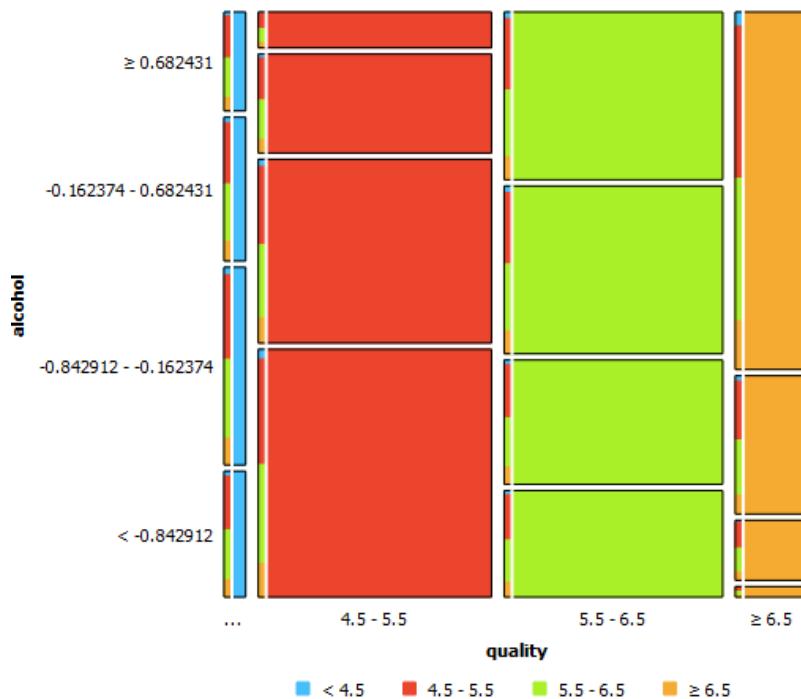
### 1.5.8. Mosaic Display

Widget Mosaic Display pozwala na wizualizację zależności między cechami a jakością wina w formie mozaiki. Każdy prostokąt reprezentuje kombinację wartości cech i jakości, a jego wielkość odpowiada liczebności tej kombinacji. Kolory prostokątów wskazują na różne poziomy jakości wina.

Ten rodzaj wizualizacji przypomina nieco wykorzystany w poprzednim podpunkcie Sieve Diagram.

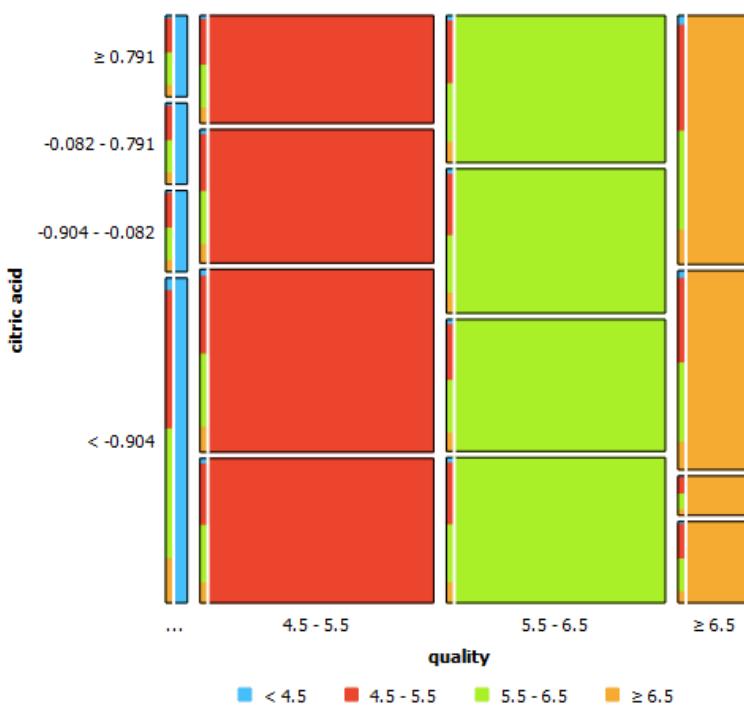
### Quality vs. Alcohol

- **Opis:** Diagram pokazuje, że wyższe wartości alkoholu ( $\geq 0.682431$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Duże prostokąty w kolorze pomarańczowym ( $\geq 6.5$ ) w tej kategorii wskazują na dużą liczbę wysokiej jakości próbek wina. Natomiast niskie wartości alkoholu ( $< -0.842912$ ) są częściej skorelowane z niższą jakością (4.5 - 5.5).
- **Wniosek:** Wyższa zawartość alkoholu jest wyraźnie skorelowana z wyższą jakością wina.



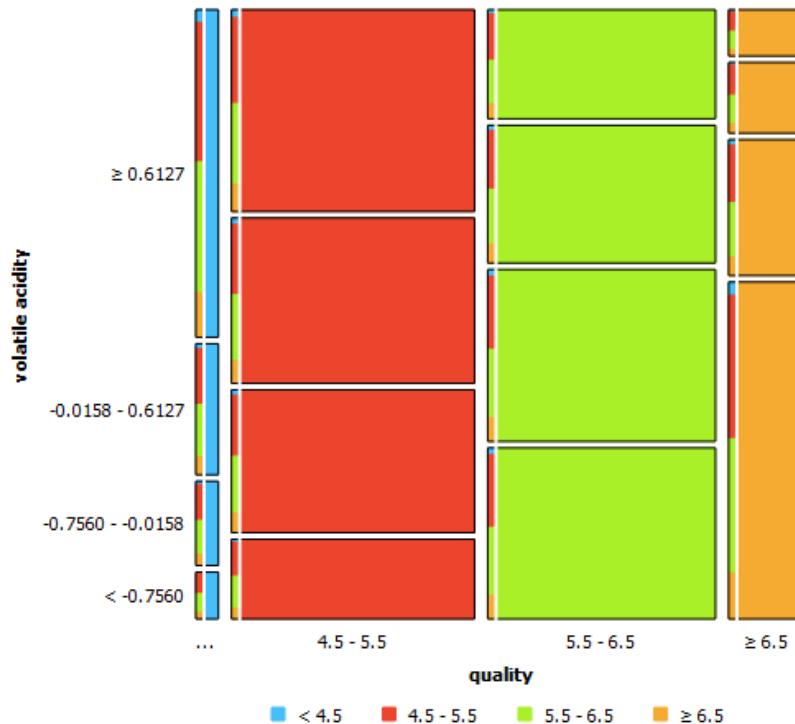
### Quality vs. Citric Acid

- Opis:** Diagram pokazuje, że wyższe wartości kwasu cytrynowego ( $\geq 0.791$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Większość próbek z niską jakością (4.5 - 5.5) ma niskie lub umiarkowane wartości kwasu cytrynowego.
- Wniosek:** Wyższa zawartość kwasu cytrynowego jest korzystna dla uzyskania wyższej jakości wina.



## Quality vs. Volatile Acidity

- **Opis:** Diagram pokazuje, że niższe wartości lotnej kwasowości ( $< -0.7560$ ) są skorelowane z wyższą jakością wina ( $\geq 6.5$ ). Próbki z wyższą lotną kwasością ( $\geq 0.6127$ ) są częściej skorelowane z niższą jakością (4.5 - 5.5).
- **Wniosek:** Niższa lotna kwasowość sprzyja wyższej jakości wina.



### 1.5.9. Podsumowanie

Na podstawie poczynionych obserwacji, możemy wywnioskować, że najbardziej istotnymi cechami są kolejno: **alcohol**, **citric acid**, **volatile acidity**, **fixed acidity**, **total sulfur dioxide**, **sulphates** i **density**.

### 4.3. Porównanie z rezultatem “feature rank”

Wyniki widgetu „Feature Rank” w dużej mierze pokrywają się z wcześniejszymi wnioskami na podstawie analiz wizualnych.

#### Zgodność z Wcześniejszymi Wnioskami:

- **Alcohol:** Jest najważniejszą cechą zarówno w analizie wizualnej, jak i według widgetu „Feature Rank”.
- **Volatile Acidity:** Również znajduje się wysoko w obu analizach, potwierdzając jej istotny wpływ na jakość wina.
- **Sulphates:** Wysokie miejsce w rankingu potwierdza wcześniejsze wnioski o ich znaczeniu.
- **Citric Acid:** Zajmuje wysoką pozycję, zgodnie z wcześniejszymi obserwacjami.
- **Total Sulfur Dioxide:** Jest istotną cechą w obu analizach.
- **Density:** Znajduje się wysoko zarówno w analizie wizualnej, jak i w wynikach widgetu „Feature Rank”.

#### Różnice:

- **Fixed Acidity:** Chociaż w analizach wizualnych była uznana za istotną, w wynikach widgetu „Feature Rank” ma niższą pozycję, co sugeruje, że jej wpływ może być mniejszy niż innych cech.
- **Chlorides:** Wysoka pozycja w wynikach widgetu „Feature Rank” wskazuje, że może mieć większy wpływ niż wskazywały to wizualizacje.

#	Univar. reg.	RReliefF
1	alcohol	468.267
2	volatile acidity	287.444
3	sulphates	107.740
4	citric acid	86.258
5	total sulfur dioxide	56.658
6	density	50.405
7	chlorides	26.986
8	fixed acidity	24.960
9	pH	5.340
10	free sulfur dioxide	4.109
11	residual sugar	0.301

## 4.4. Które cechy są najbardziej zbliżone

Aby określić, które cechy są najbardziej zbliżone, możemy skorzystać z wcześniejszych analiz wizualnych, takich jak Heat Map i FreeViz, które pokazują wzorce współzmienności i klasteryzacji cech.

### Najbardziej Zbliżone Cechy:

#### 1. Free Sulfur Dioxide i Total Sulfur Dioxide:

- Te cechy są ze sobą ścisłe powiązane, ponieważ obie dotyczą zawartości dwutlenku siarki w winie.

#### 2. Citric Acid i Fixed Acidity:

- Obie cechy dotyczą kwasowości wina i często wykazują podobne wzorce wartości.

#### 3. Chlorides i Density:

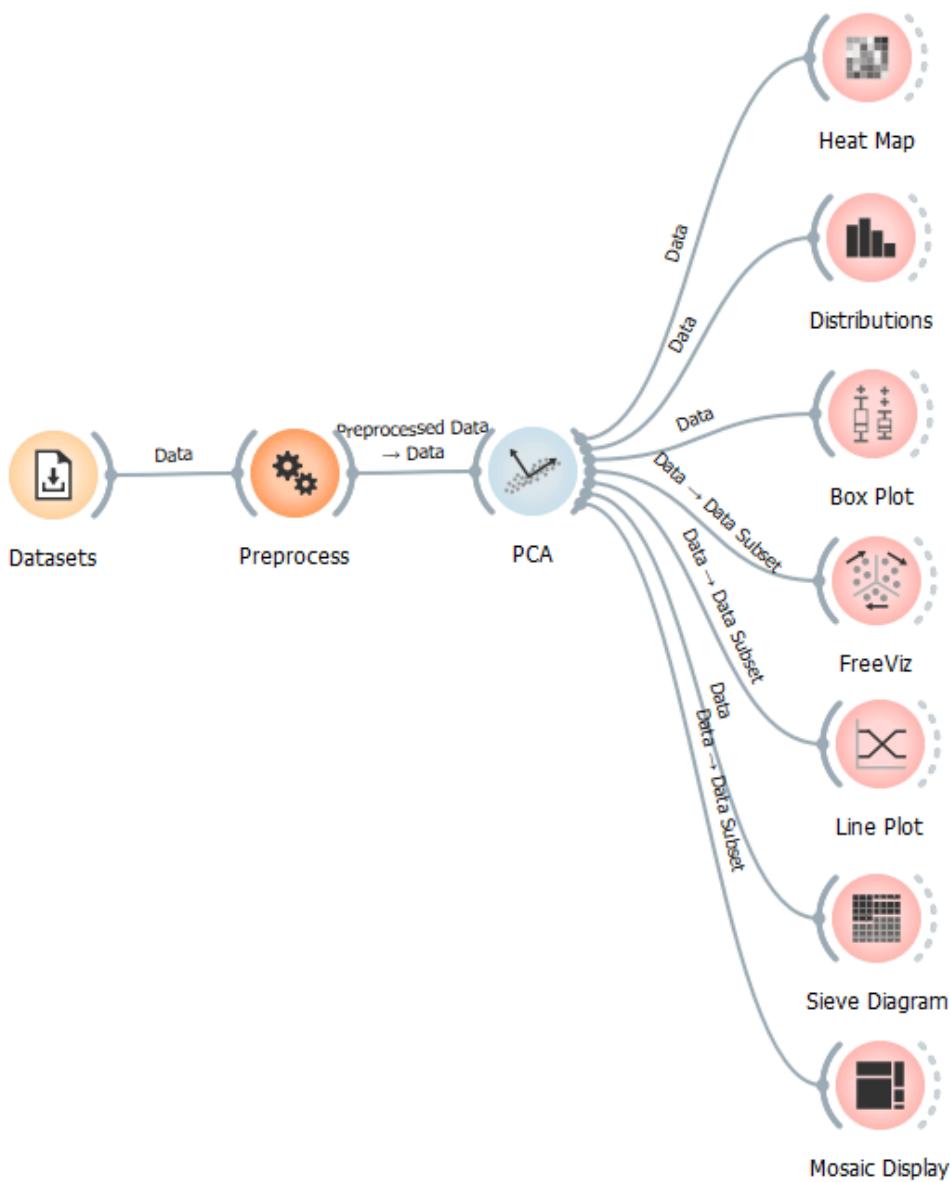
- Wzorce wartości tych cech są zbliżone, co sugeruje ich wzajemny wpływ na siebie.

#### 4. Volatile Acidity i pH:

- Te cechy również pokazują pewne podobieństwa w rozkładzie wartości, wskazując na powiązanie między nimi.

## 4.5. Wizualizacja z wykorzystaniem PCA

Poniżej zamieściłem diagram, który zostanie wykorzystany przy analizie z wykorzystaniem PCA.



### 3.1.1. Heat Map

#### Najważniejsze Cechy:

- PC1: Największy wpływ mają **volatile acidity**, **citric acid**, **fixed acidity**, **density**, **alcohol**, **total sulfur dioxide**, i **sulphates**.
- PC2: Kluczowe są **chlorides**, **free sulfur dioxide**, i **residual sugar**.

### Grupowanie:

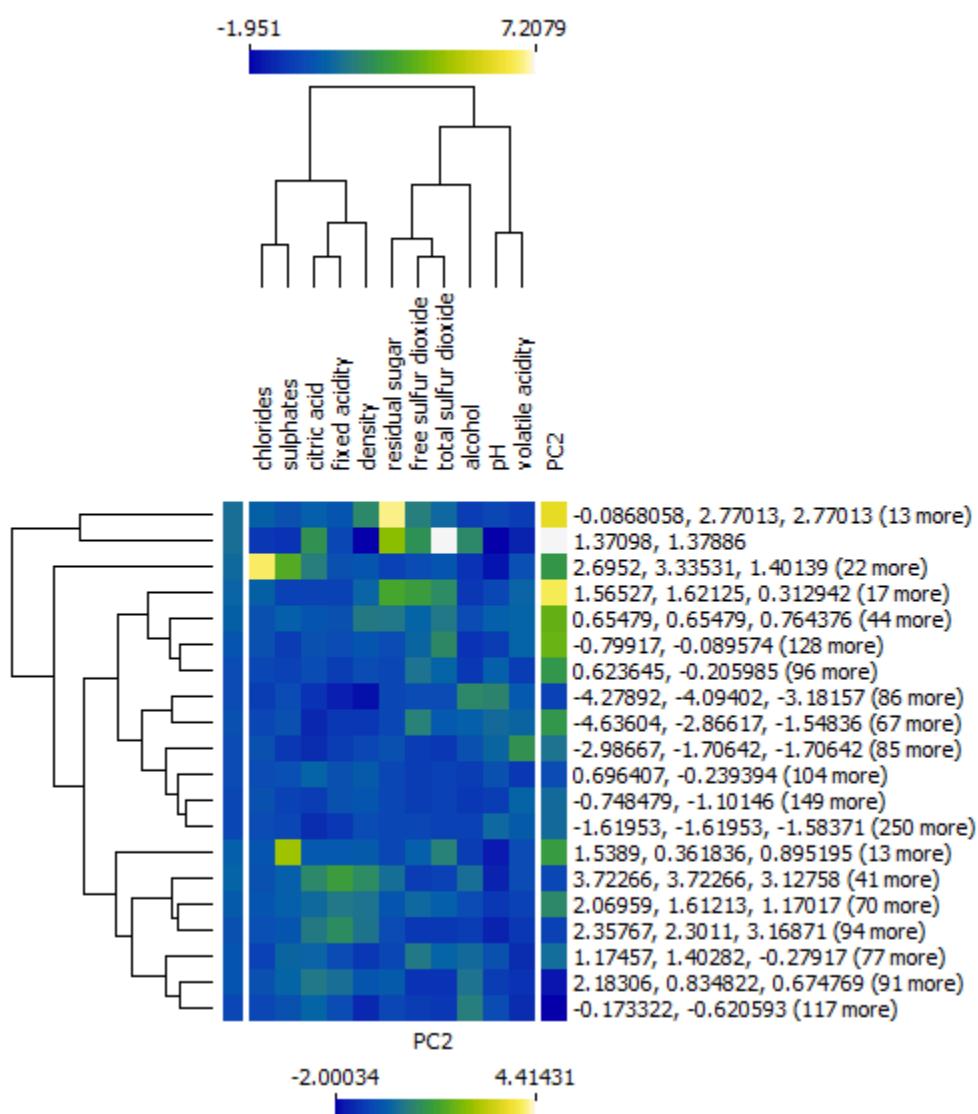
- Cechy **volatile acidity**, **citric acid**, **fixed acidity**, **density**, i **alcohol** grupują się razem, potwierdzając ich istotność.
- **Chlorides**, **free sulfur dioxide**, i **residual sugar** mają większy wpływ na PC2, co może sugerować ich dodatkowe znaczenie.

### Porównanie z Poprzednią Analizą:

- **Podobieństwa:** **Alcohol**, **volatile acidity**, **citric acid**, **total sulfur dioxide**, i **sulphates** są nadal kluczowe.
- **Różnice:** **Chlorides**, **free sulfur dioxide**, i **residual sugar** wykazują większy wpływ w PCA.

### Wnioski

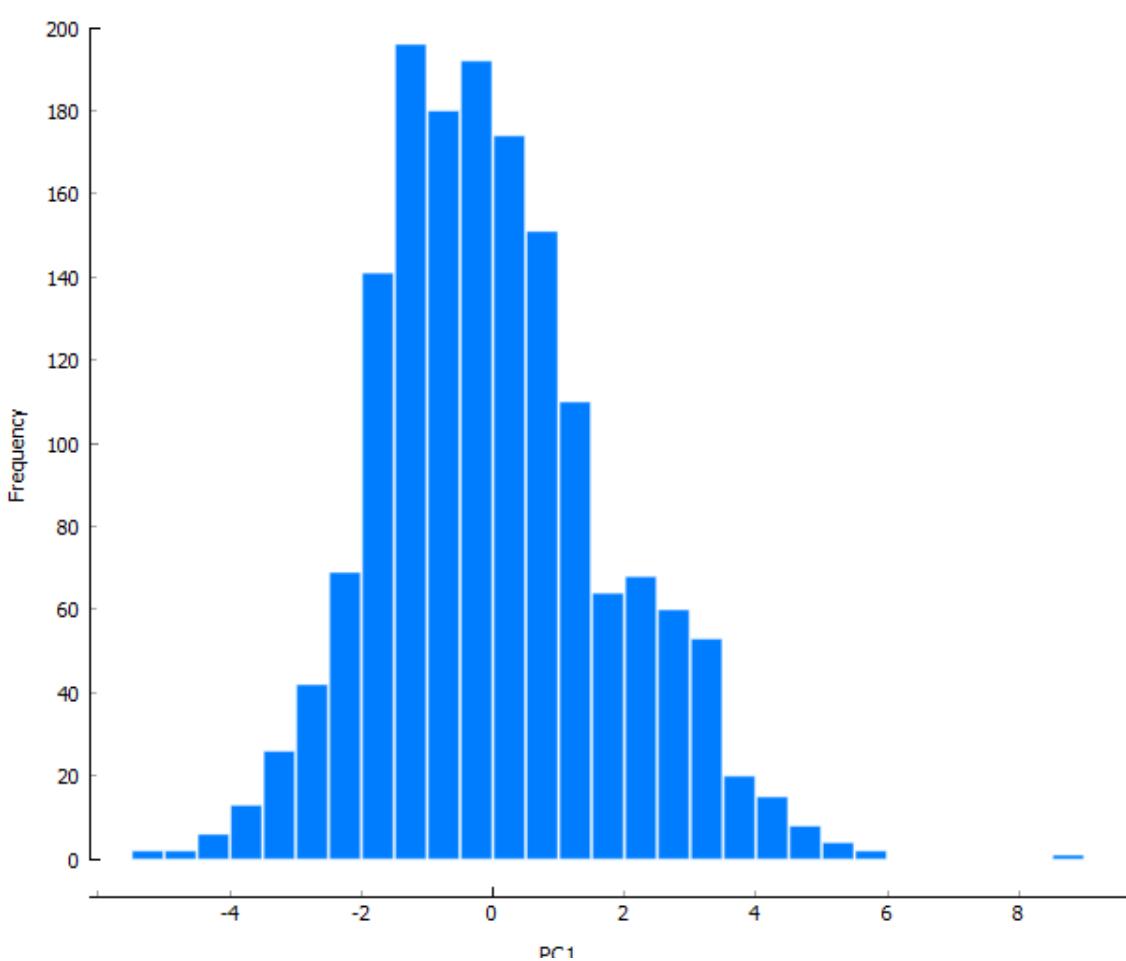
Analiza PCA potwierdza wcześniejsze wnioski i uwidacznia dodatkowe znaczenie niektórych cech, takich jak **chlorides** i **free sulfur dioxide**.



### 3.1.2. Distributions

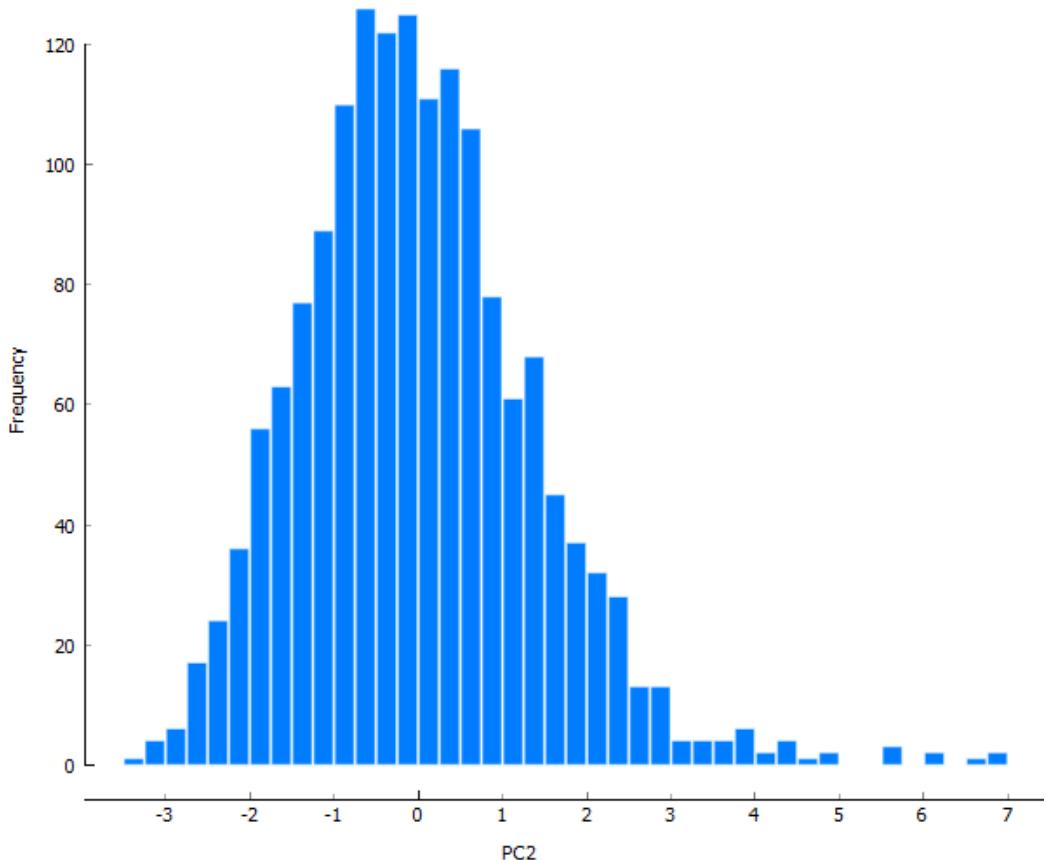
#### Rozkład PC1:

- **Opis:** Histogram przedstawia rozkład wartości dla pierwszej głównej składowej (PC1). Wartości PC1 są rozłożone symetrycznie wokół wartości 0, co sugeruje normalny rozkład.
- **Wnioski:** PC1 wyjaśnia znaczącą część wariancji w danych. Wartości PC1 skupiają się głównie wokół środka rozkładu, co wskazuje na centralną tendencję cech składających się na tę składową.



#### Rozkład PC2:

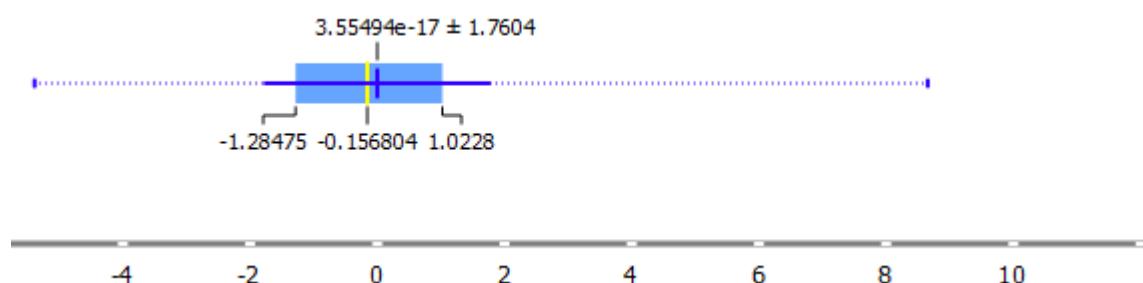
- **Opis:** Histogram przedstawia rozkład wartości dla drugiej głównej składowej (PC2). Wartości PC2 również wykazują rozkład zbliżony do normalnego, choć z nieco większą asymetrią po prawej stronie.
- **Wnioski:** PC2 wyjaśnia dodatkową część wariancji w danych. Rozkład PC2 jest bardziej asymetryczny, co może sugerować większą różnorodność w cechach składających się na tę składową.



### 3.1.3. Box Plot

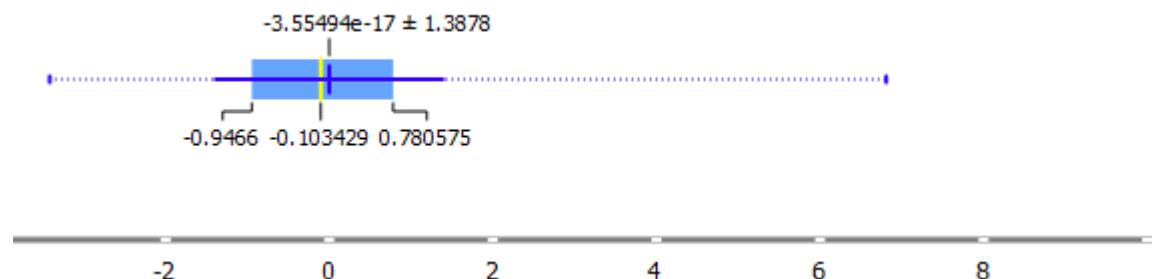
#### Box Plot dla PC1

- **Średnia:** Wartość średnia (0) jest bardzo blisko zera, co wskazuje na centralną tendencję rozkładu PC1.
- **Rozstęp międzykwartylowy (IQR):** Większość danych znajduje się w przedziale od około -1.28 do 1.02, co pokazuje, że rozkład wartości PC1 jest dość wąski i skupiony wokół środka.
- **Wartości skrajne:** Obserwujemy kilka wartości skrajnych po obu stronach rozkładu, ale większość danych mieści się w wąskim przedziale.



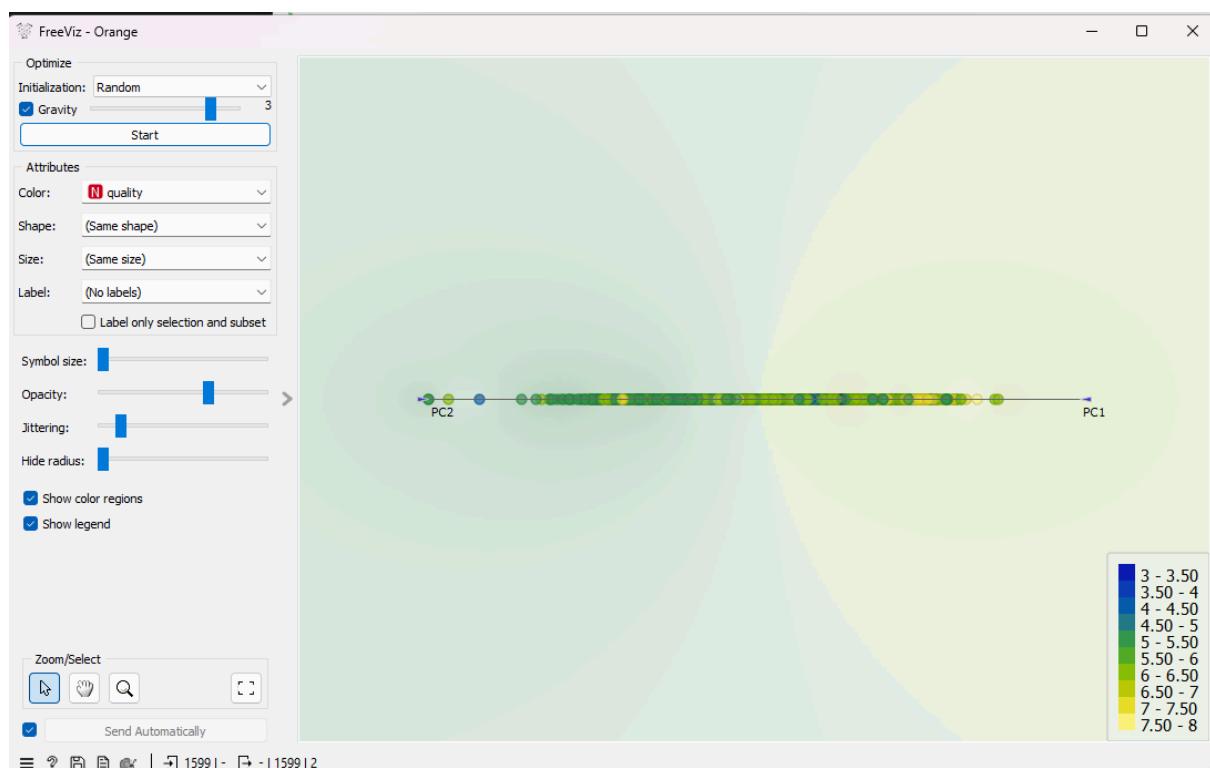
### Box Plot dla PC2

- **Średnia:** Podobnie jak w przypadku PC1, średnia dla PC2 jest bardzo blisko zera.
- **Rozstęp międzykwartylowy (IQR):** Większość danych znajduje się w przedziale od około -0.95 do 0.78, co sugeruje, że wartości PC2 również są skupione wokół środka, ale z mniejszą zmiennością niż PC1.
- **Wartości skrajne:** Widać kilka wartości skrajnych, jednak rozkład jest bardziej skupiony niż w przypadku PC1.



### 3.1.4. FreeViz

Otrzymane wyniki dla widgetu FreeViz przedstawiają dane w postaci dwóch głównych składowych (PC1 i PC2), pokazując rozmieszczenie próbek wina według ich jakości (**quality**).



### **Wnioski:**

- Wizualizacja pokazuje, że jakość wina jest w miarę równomiernie rozłożona wzdłuż głównych składowych, z tendencją do wyższych jakości na końcach wykresu. Można zauważać, że wina o wyższej jakości (kolory żółty i zielony) są skoncentrowane bardziej na prawym końcu PC1.
- PC1 i PC2 skutecznie redukują wielowymiarowość danych, zachowując najważniejsze informacje. Rozkład punktów sugeruje, że obie składowe mają znaczący wpływ na zmienność w danych.

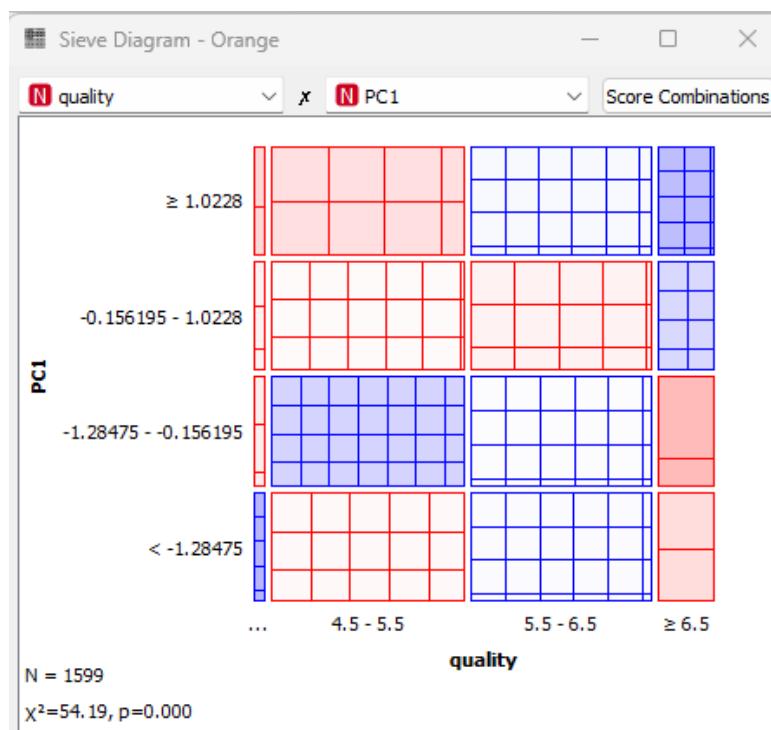
### **3.1.5. Line Plot**

Ponownie pomijamy, ponieważ nie ma z czym porównywać, jak wcześniej został pominięty ten widget w analizie.

### **3.1.6. Sieve Diagram**

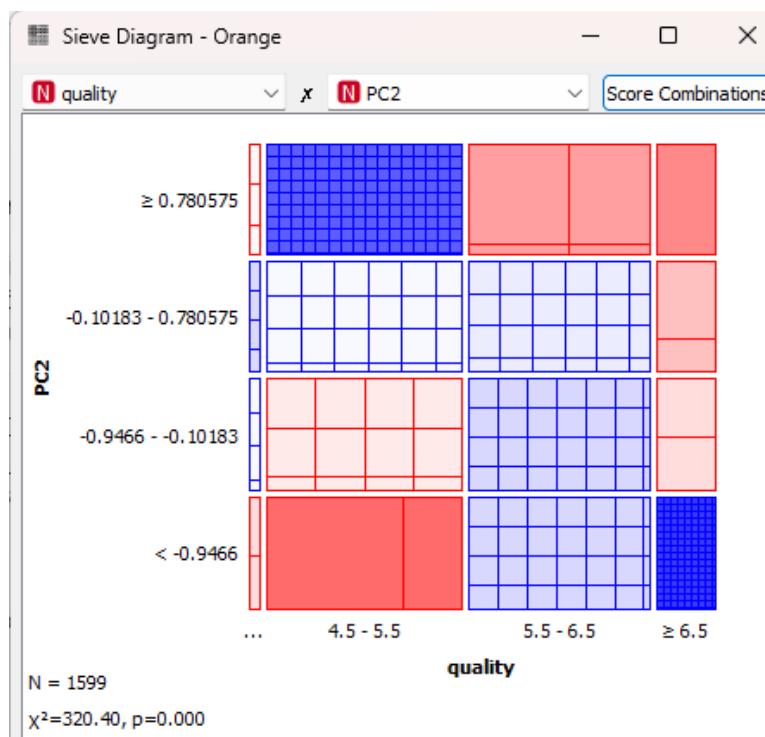
#### **PC1 vs Quality**

- Wysokie wartości PC1 ( $\geq 1.0228$ ) mają tendencję do łączenia się z wysoką jakością wina ( $\geq 6.5$ ), co sugeruje, że większe wartości PC1 są skorelowane z lepszą jakością.
- Niskie wartości PC1 ( $< -1.28475$ ) są związane głównie z niższą jakością ( $< 4.5$ ), co sugeruje, że niższe wartości PC1 są skorelowane z gorszą jakością wina.
- Średnie wartości PC1 (-1.28475 do 1.0228) pokazują rozproszenie jakości od 4.5 do 6.5, co wskazuje na bardziej zmieszane relacje w tym zakresie.



## PC2 vs Quality

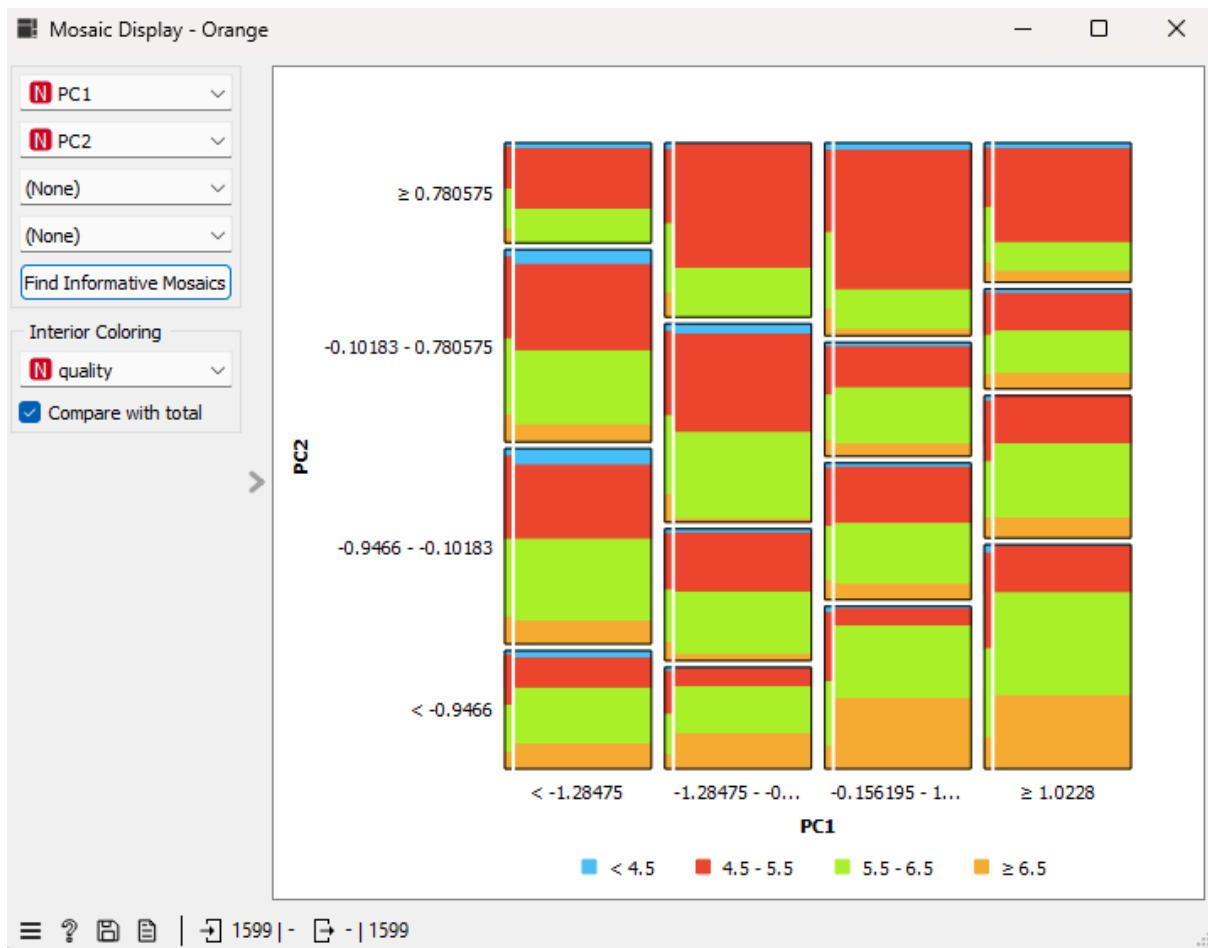
- Wysokie wartości PC2 ( $\geq 0.780575$ ) są skorelowane z niską jakością wina (4.5 - 5.5), co sugeruje, że wyższe wartości PC2 mogą być związane z niższą jakością.
- Niskie wartości PC2 ( $< -0.9466$ ) są związane z lepszą jakością wina ( $\geq 6.5$ ), co sugeruje, że niższe wartości PC2 są skorelowane z lepszą jakością wina.
- Średnie wartości PC2 (-0.9466 do 0.780575) wykazują bardziej zmieszane relacje z jakością wina, podobnie jak w przypadku PC1.



### 3.1.7. Mosaic Display

Ten rodzaj wizualizacji pozwala łatwiej zaobserwować to, co dało się wywnioskować z obserwacji wyżej umieszczonych diagramów (Sieve Diagram).

Na poniższym diagramie widać, że im jednocześnie wartości PC1 są większe i PC2 mniejsze (prawy dolny róg), tym więcej jest próbek dobrej jakości. Jakość spada wraz ze zmniejszaniem się wartości PC1 oraz ze wzrostem wartości PC2.



### 3.1.8. Podsumowanie

Analiza PCA (Principal Component Analysis) pozwoliła zredukować złożoność danych, identyfikując główne składowe, które mają największy wpływ na zmienność w zbiorze danych. Na podstawie wizualizacji i analizy głównych składowych (PC1 i PC2), wyciągnięto następujące wnioski:

1. **Wysokie wartości PC1 w połączeniu z niskimi wartościami PC2** są związane z **wyższą jakością wina**.
2. **Niskie wartości PC1 i wysokie wartości PC2** są skorelowane z **niższą jakością wina**.
3. **Ekstremalne wartości** głównych składowych mają **istotny wpływ na jakość wina**, podczas gdy średnie wartości wykazują bardziej zróżnicowane wyniki.

PCA potwierdziło wcześniejsze obserwacje z analiz wizualnych, wskazując na złożoność relacji między cechami chemicznymi wina a jego jakością.

## 4.6. Wnioski na podstawie wizualizacji Heat Map oraz Distance Table

Analiza heat map i distance table dostarcza dodatkowych informacji na temat relacji między cechami wina. Oto wnioski, które można wyciągnąć na podstawie przeprowadzonej analizy:

### **Heat Map:**

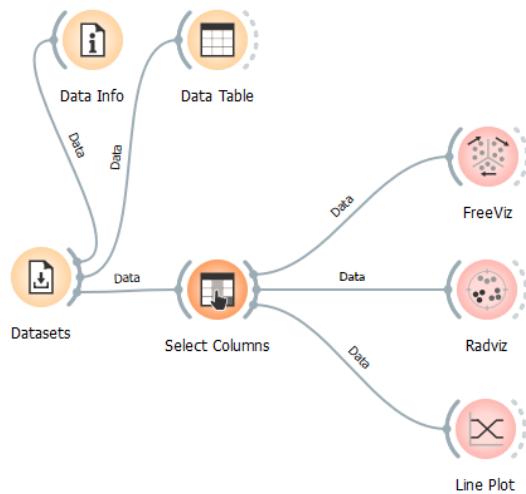
- **Grupowanie cech:** Cechy takie jak chlorides, sulphates i citric acid oraz fixed acidity, density i pH grupują się blisko siebie, co sugeruje, że mają podobny wpływ na jakość wina.
- **Ekstremalne wartości:** Wartości cech, które są skrajnie wysokie lub niskie, wykazują silniejszy wpływ na jakość wina, co jest widoczne w kontrastach kolorystycznych na mapie ciepła.

### **Distance Table:**

- **Podobieństwo cech:** Cechy blisko siebie w tabeli odległości wykazują wysokie podobieństwo, co może wskazywać na redundantność informacji.
- **Różnorodność wpływów:** Tabela odległości pomaga zidentyfikować cechy o unikalnym wpływie na jakość wina, które nie są ściśle skorelowane z innymi cechami.

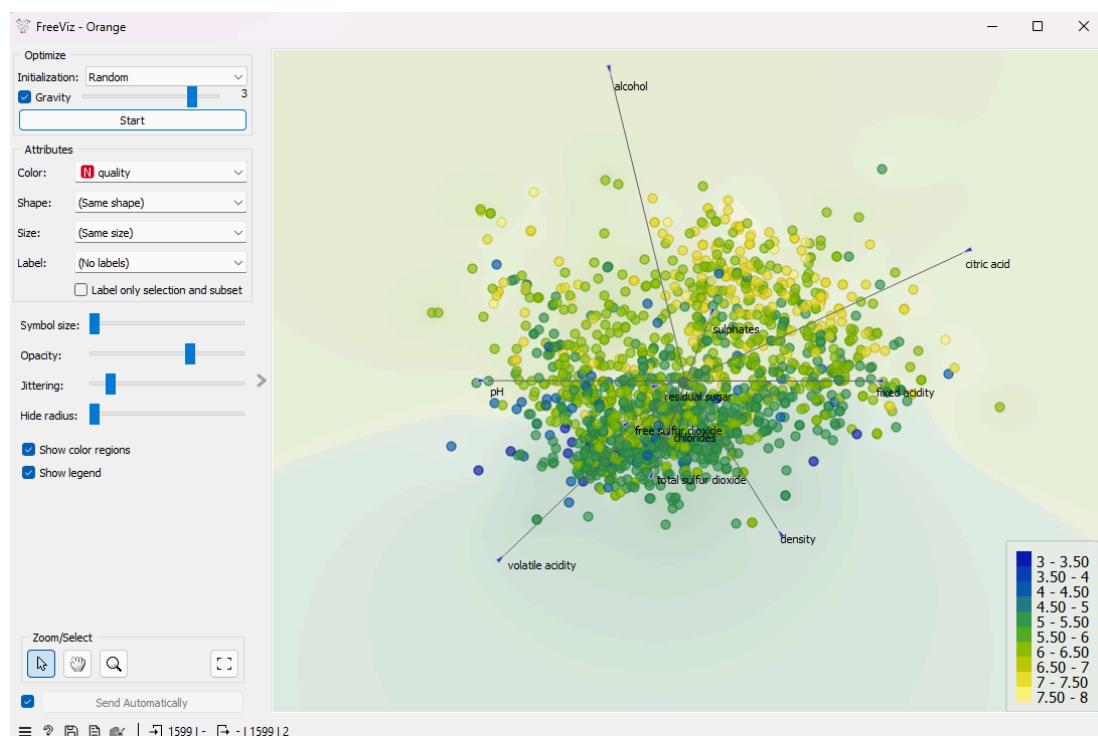
## 2. Zadanie 2

W tym zadaniu wykorzystałem zamieszczony na poniższej ilustracji diagram, służący do analizy cech, przy pomocy widgetów, takich jak: FreeViz, RadViz oraz Line Plot.



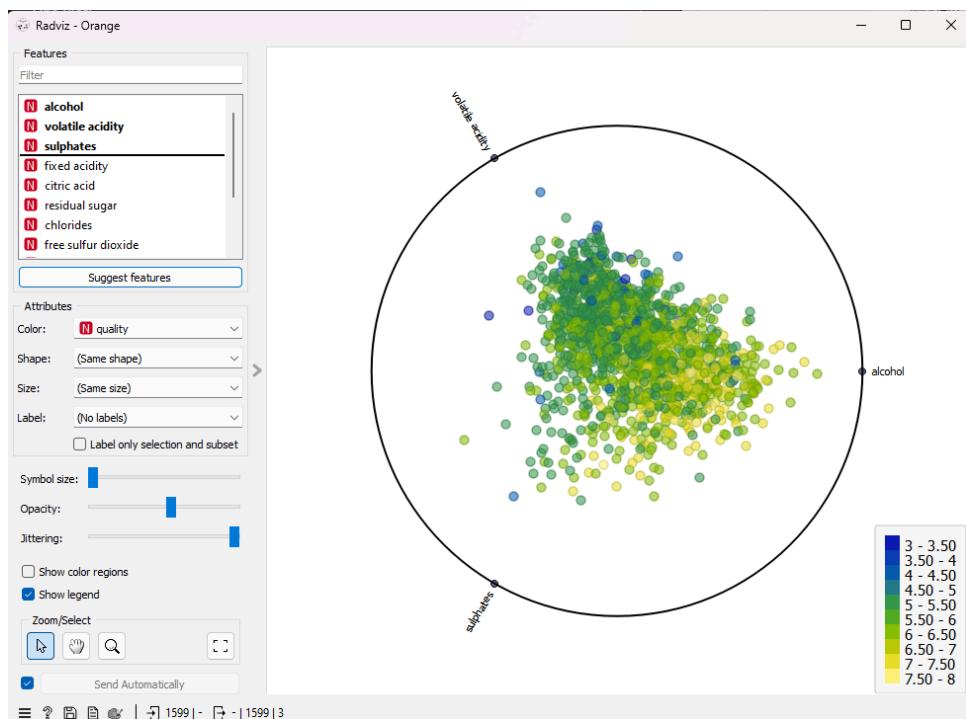
### 4.1. FreeViz

- Wizualizacja pokazuje, że cechy takie jak alkohol, lotna kwasowość i siarczany mają znaczący wpływ na jakość wina. Próbki o wyższej jakości (oznaczone kolorem żółtym) są bardziej skoncentrowane w kierunku wektorów dla tych cech.
- Widać wyraźne skupienie próbek o niskiej jakości (kolor niebieski) w przeciwnym kierunku.

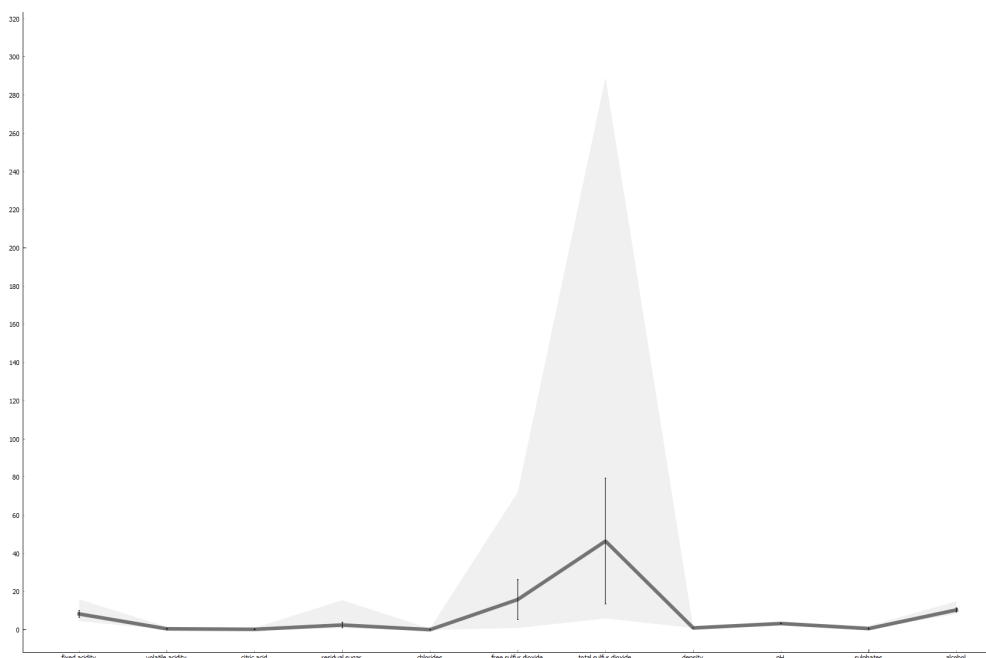


## 4.2. RadViz

- W przypadku wizualizacji RadViz wybrałem najbardziej istotne cechy, wybrane przy pomocy wizualizacji FreeViz. Dzięki tej wizualizacji, w łatwy sposób możemy uszeregować najważniejsze cechy według istotności.
- Widzimy więc, że spośród trzech najistotniejszych cech, największy wpływ na jakość wina ma cecha alcohol.



## 4.3. Line Plot



## 4.4. Podsumowanie

Wszystkie trzy wizualizacje potwierdzają wcześniejsze wnioski z heat mapy i tabeli odległości, wskazując, że alkohol, lotna kwasowość i siarczany są kluczowymi cechami wpływającymi na jakość wina. Próbki o wyższej jakości mają wyższe wartości tych cech, podczas gdy próbki o niższej jakości mają wartości niższe.

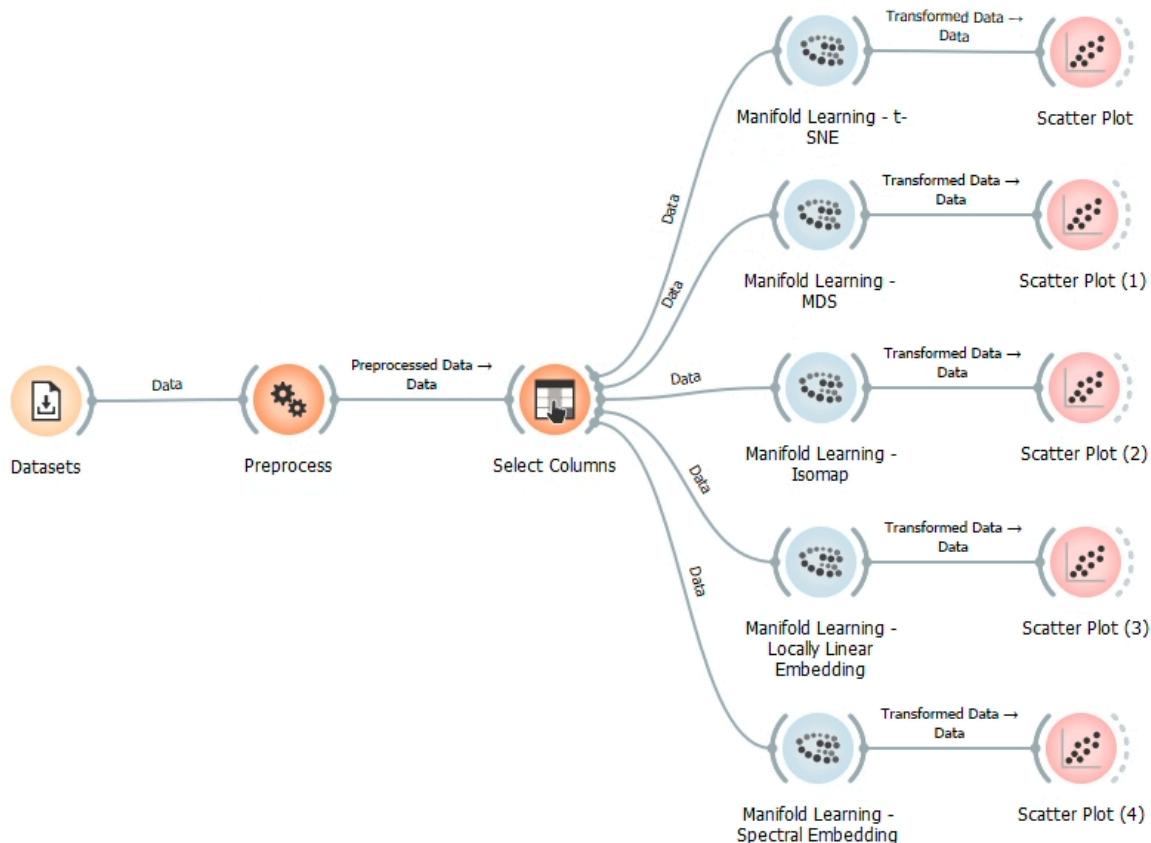
### 3. Zadanie 3

#### 4.1. Najlepszy algorytm embeddingu oraz porównania wizualizacji

##### 3.1.1. Poszukiwanie najlepszego algorytmu embeddingu

W pierwszym kroku zajmę się znalezieniem najlepszego algorytmu embeddingu. Do tego celu wykorzystam poniższy graf, w którym najpierw standaryzuję dane do  $\mu = 0$  i  $\sigma^2=1$ , a następnie, wykorzystuję kolejno wymienione algorytmy embeddingu, które wizualizuję w postaci Scatter Plotów:

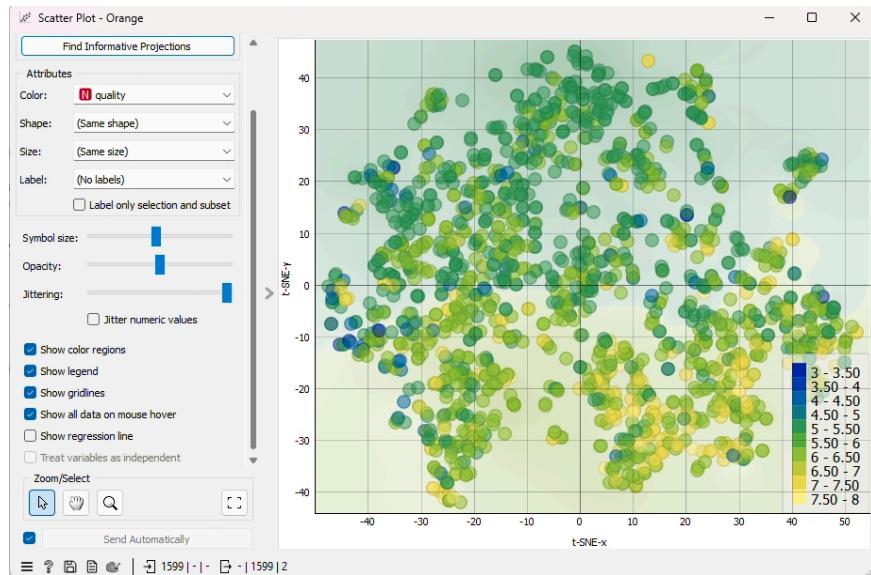
- t-SNE,
- MDS,
- Isomap,
- Locally Linear Embedding,
- Spectral Embedding



## t-SNE

**Opis:** t-SNE pokazuje dość równomiernie rozłożone punkty, co może sugerować dobrą separację klastrów. Widoczna jest kolorowa różnorodność w różnych regionach wykresu, co sugeruje, że t-SNE dobrze oddziela różne klasy.

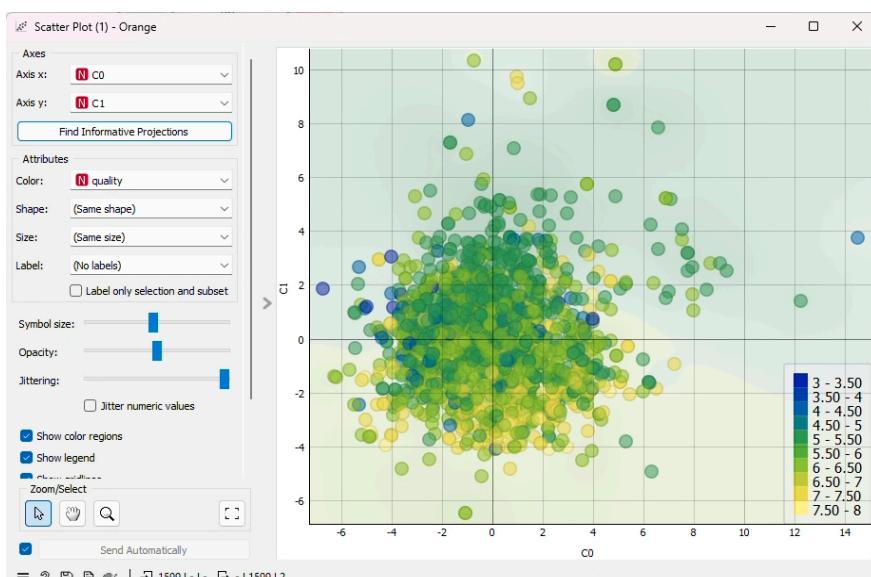
**Ocena:** t-SNE jest często skuteczny w wizualizacji struktury danych w niższej wymiarowości, zwłaszcza gdy dane mają złożone nielinowe zależności.



## MDS (Multidimensional Scaling)

**Opis:** Wykres z MDS wydaje się mieć mniej wyraźnie oddzielone klastry niż t-SNE. Punkty są bardziej skoncentrowane w jednym obszarze, co może sugerować trudności w wizualizacji separacji klas.

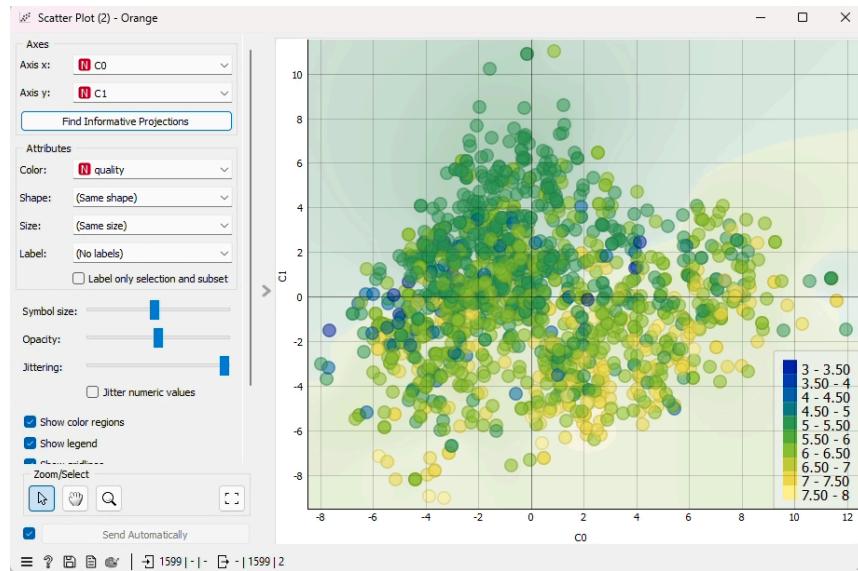
**Ocena:** MDS może być mniej efektywny niż t-SNE w przypadku danych o złożonej strukturze.



## Isomap

**Opis:** Isomap pokazuje pewne separacje między punktami, ale klastry nie są tak wyraźnie widoczne jak w przypadku t-SNE. Widać pewne zagęszczenie punktów, co może wskazywać na ograniczoną skuteczność w przypadku bardziej złożonych danych.

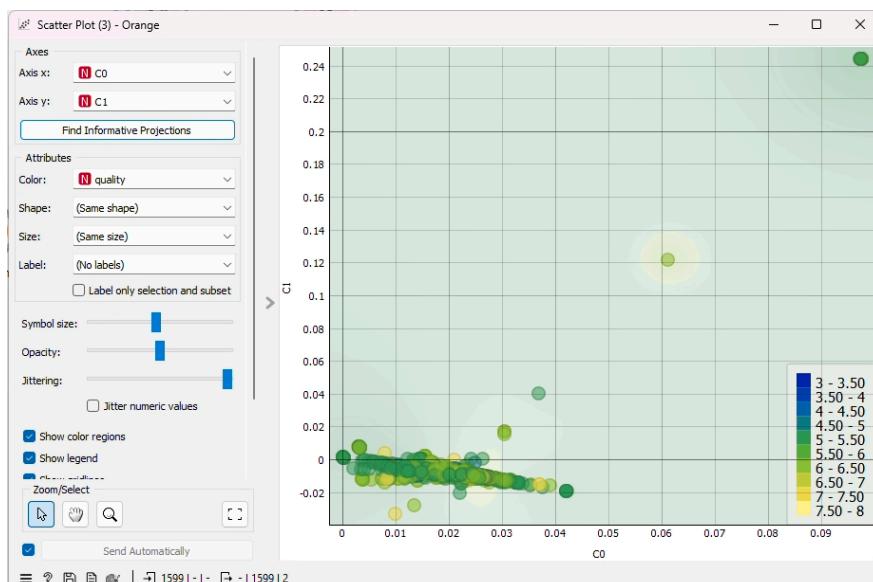
**Ocena:** Isomap może być skuteczny dla danych o bardziej liniowej strukturze, ale w przypadku bardziej złożonych danych może mieć ograniczenia.



## Locally Linear Embedding (LLE)

**Opis:** Wykres z LLE pokazuje pewną separację, ale ogólnie punkty wydają się być mocno skoncentrowane w jednym obszarze. Brakuje wyraźnych klastrów, co może sugerować, że LLE nie jest najlepszym wyborem dla tych danych.

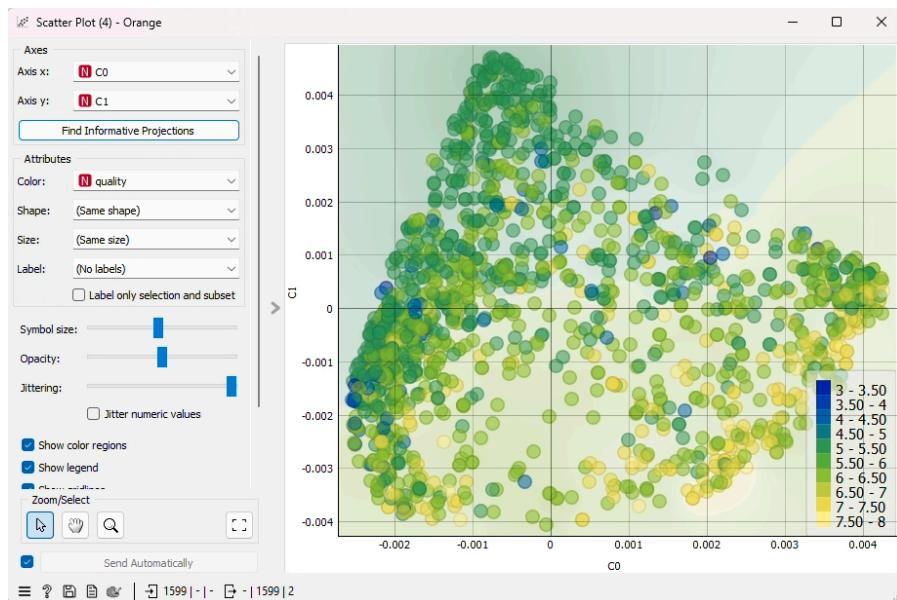
**Ocena:** LLE może mieć trudności z zachowaniem globalnej struktury danych, co może prowadzić do mniej skutecznych wizualizacji.



## Spectral Embedding

**Opis:** Wykres z Spectral Embedding pokazuje dość jednolitą koncentrację punktów, bez wyraźnych klastrów. Podobnie jak LLE, może mieć trudności z uchwyceniem bardziej złożonej struktury danych.

**Ocena:** Spectral Embedding może być mniej skuteczny niż t-SNE w wizualizacji danych o złożonych zależnościach nieliniowych.



### 3.1.2. Wybór najlepszego algorytmu embeddingu

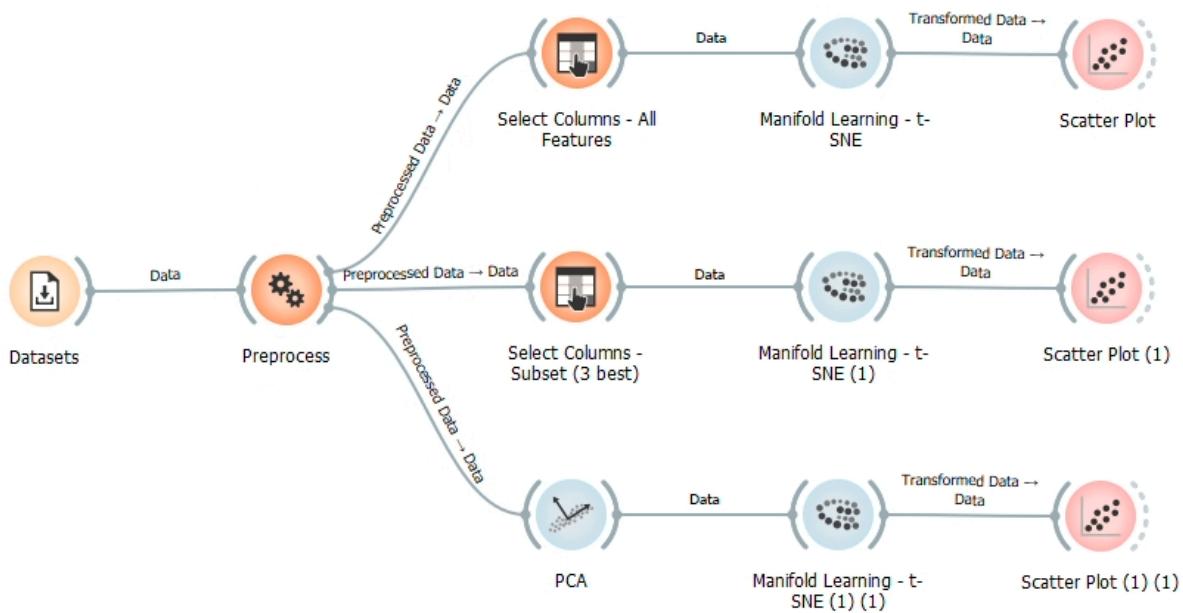
Na podstawie wizualnej analizy wykresów, **t-SNE** wydaje się być najlepszym algorytmem embeddingu dla danych w analizowanym zbiorze. t-SNE dobrze oddziela różne klastry i prezentuje wyraźną strukturę danych w dwuwymiarowej przestrzeni. Pozostałe metody, takie jak MDS, Isomap, LLE, i Spectral Embedding, wykazują pewne ograniczenia w wizualizacji separacji klas i ogólnej struktury danych.

### 3.1.3. Porównanie wizualizacji dla całego zbioru oraz podzbioru 3 najlepszych cech/głównych składowych

Do porównania wizualizacji wykorzystałem widoczny na następnej stronie opracowania graf w Orange.

W przypadku wyboru podzbioru najlepszych cech, cechami, jakie wybrałem, są: **alcohol**, **volatile acidity** oraz **sulphates**, czyli cechy, które okazały się najbardziej wpływowe na jakość wina, zgodnie z przeprowadzoną wcześniej analizą.

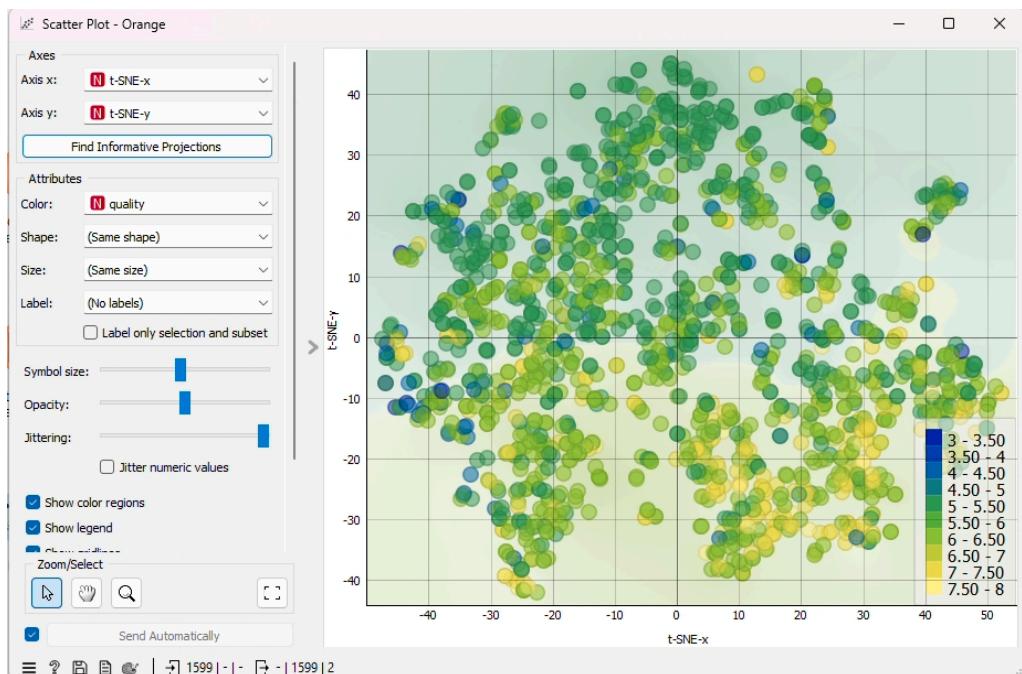
W przypadku PCA, zgodnie z poleceniem, ustawiłem liczbę głównych składowych (PC) na 3, czylili tyle samo, co liczba najlepszych cech przy wyborze cech.



## Wszystkie cechy

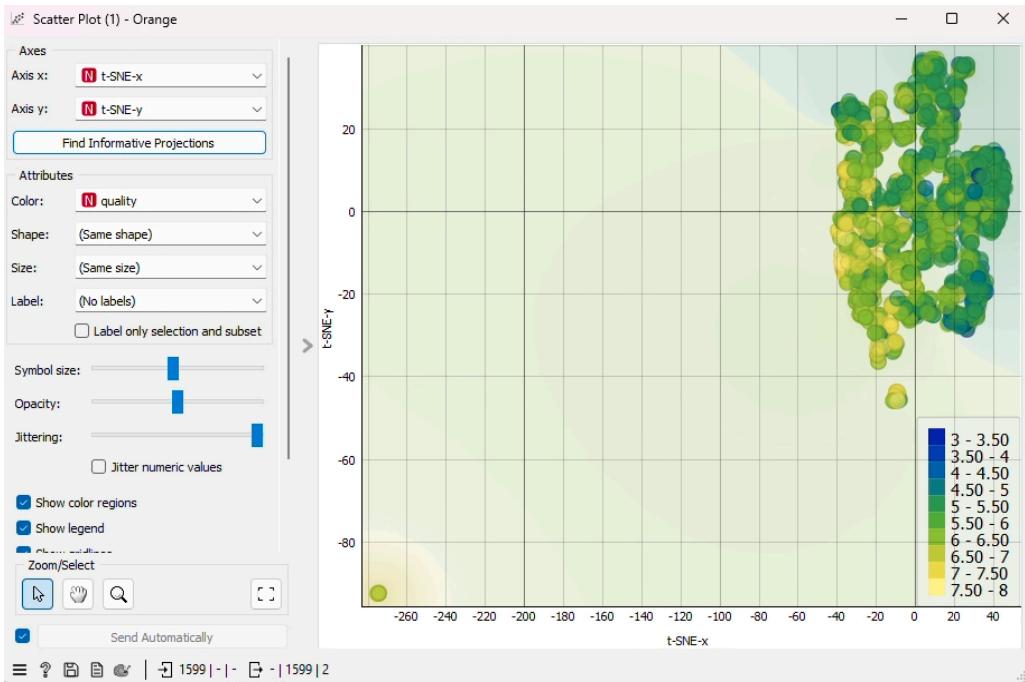
**Opis:** Wykres pokazuje rozłożone punkty na całym obszarze, z widocznymi różnicami w kolorach, co sugeruje pewną separację klas. Kolory są dość równomiernie rozmieszczone, co może oznaczać, że dane są dobrze reprezentowane przez pełny zestaw cech.

**Ocena:** Wydaje się, że pełny zestaw cech pozwala na dość dobrą separację danych, jednak nie widać wyraźnych klastrów.



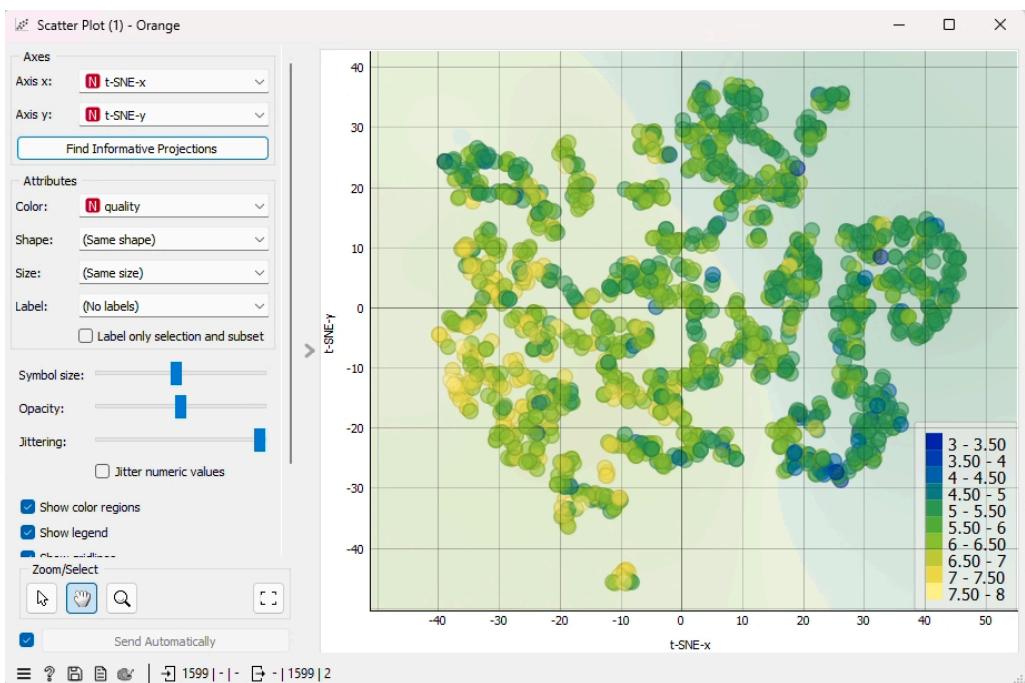
## 3 najlepsze cechy

**Opis:** Widzimy, że jeden skrajny punkt, leżący w lewym dolnym rogu wykresu, nieco zaciemnia wizualizację. Z tego powodu, odrzucimy ten punkt i przyjrzymy się bliżej pozostałojej części.



**Opis po odrzuceniu skrajnego punktu:** Wykres po odrzuceniu punktu skrajnego pokazuje bardziej równomierne rozłożenie punktów z wyraźnymi klastrami. Separacja klas jest bardziej wyraźna, a kolorowanie pokazuje lepszą strukturę danych.

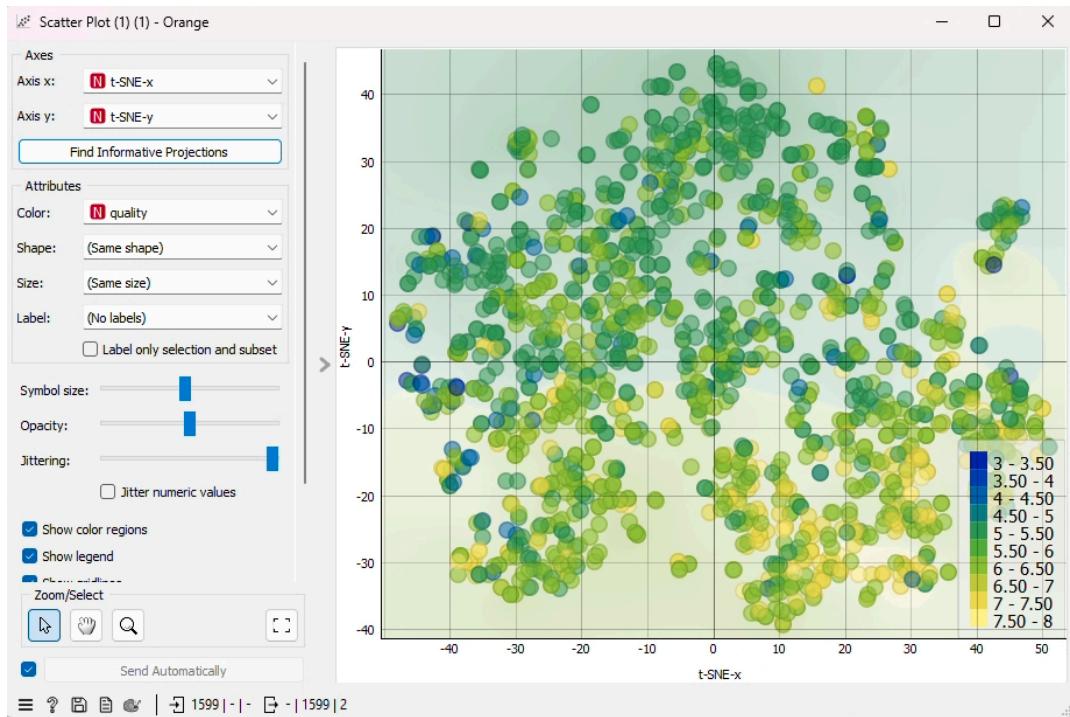
**Ocena:** Wizualizacja jest znacznie bardziej czytelna niż poprzednia wersja z trzema cechami. Separacja klas jest wyraźna, co sugeruje, że te trzy najbardziej istotne cechy mogą być wystarczające do reprezentacji zbioru danych.



### 3 główne składowe (PC)

**Opis:** Wykres pokazuje punkty rozłożone podobnie jak w przypadku pełnego zestawu cech, z lepszą widocznością niektórych klastrów. Kolory są dobrze rozdzielone, co sugeruje, że PCA dobrze oddaje strukturę danych w trzech głównych składowych.

**Ocena:** Dobra separacja klas i czytelna wizualizacja. PCA zapewnia dobre odwzorowanie danych.



## 4.2. Wybór najlepszych cech, czy głównych składowych daje lepszą wizualizację?

Porównując otrzymane w poprzednim punkcie wykresy, można wywnioskować, że lepszą wizualizację otrzymujemy dla 3 najistotniejszych cech niż dla 3 głównych składowych. Spowodowane jest to tym, że na wykresie dla 3 najbardziej istotnych cech są lepiej widoczne klastry, zależne od jakości wina. Wykres ten jest bardziej czytelny, pozwalając na łatwiejszą interpretację danych.

## 4.3. Analiza wartości odstających

### Pełny zestaw cech

Wykres pokazuje równomierne rozłożenie punktów bez wyraźnych anomalii. Pomimo dobrej separacji kolorów, brak wyraźnych klastrów wskazuje, że dane mogą zawierać szum lub redundancję cech, co utrudnia identyfikację wartości odstających.

### Podzbiór 3 najlepszych cech (po odrzuceniu punktu skrajnego)

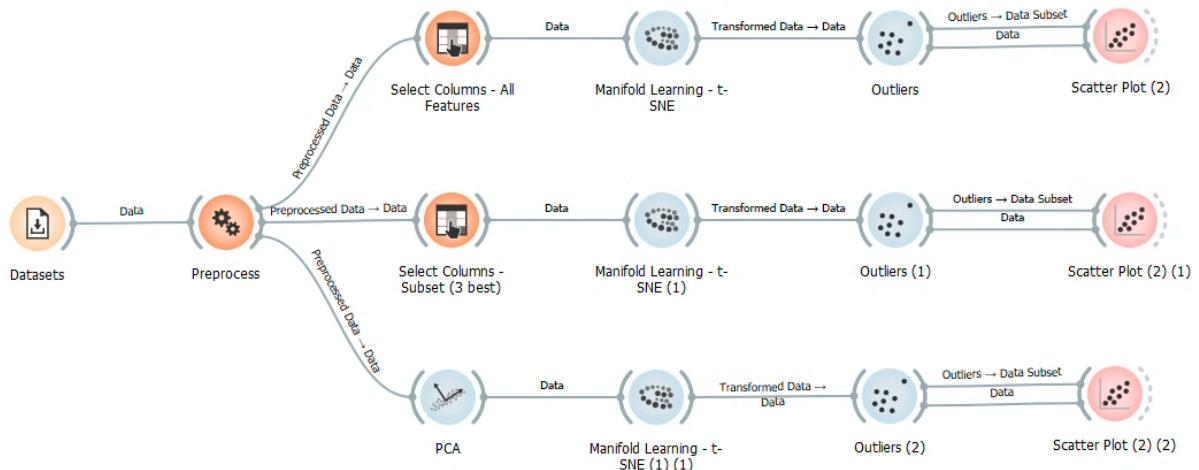
Po odrzuceniu punktu skrajnego w lewym dolnym rogu, wykres pokazuje bardziej klarowną strukturę danych z wyraźnymi klastrami. Usunięcie odstającego punktu poprawiło czytelność wizualizacji, co pozwala na lepsze zrozumienie struktury danych. Pozostałe dane nie wydają się zawierać wyraźnych anomalii.

### Trzy główne składowe (PC)

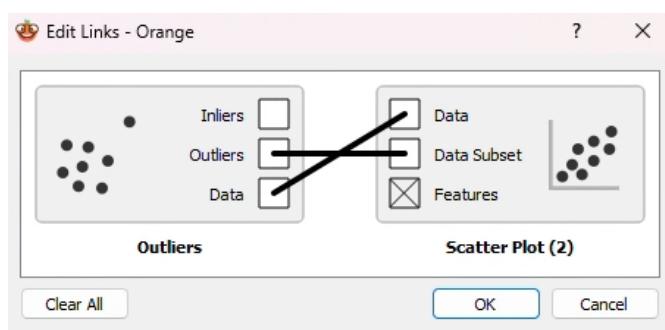
Wykres pokazuje dane rozłożone podobnie do pełnego zestawu cech, ale z wyraźniejszymi klastrami. Wartości odstające są mniej widoczne niż w pełnym zestawie cech, co sugeruje, że PCA pomogło w redukcji szumu i lepszym odwzorowaniu struktury danych.

## 4.4. Wizualizacje wartości odstających

Do wizualizacji wartości odstających wykorzystałem poniższy graf. Między węzły Manifold Learning a Scatter Plot wstawiałem węzły, które wyznaczają wartości odstające (Outliers).



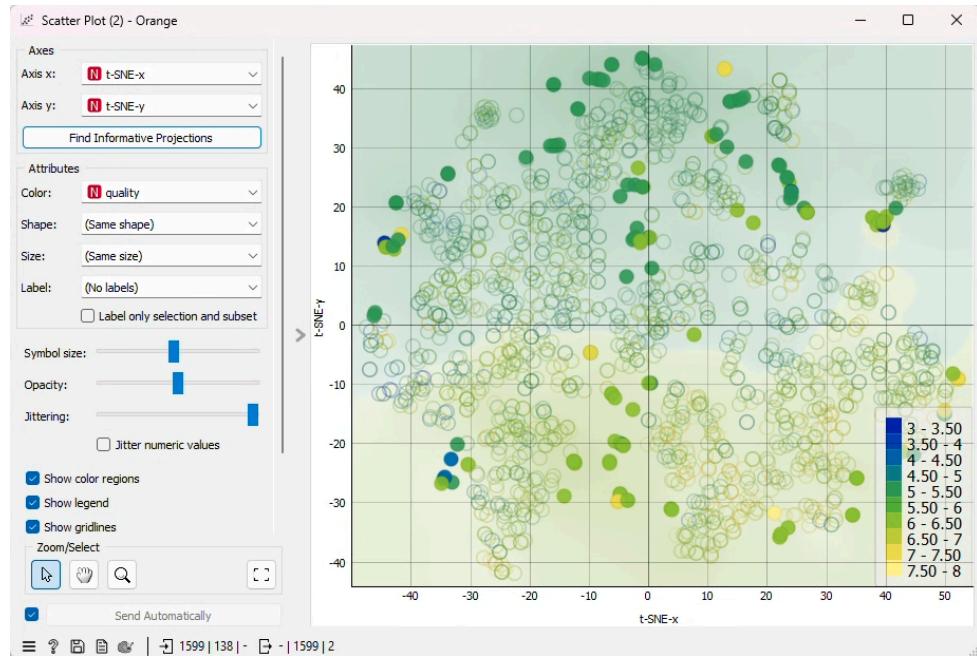
Dla każdej pary Outliers → Scatter Plot, wybrałem poniższe połączenia. Dzięki temu, widoczny jest pełen zbiór danych oraz wyszczególnione wartości odstające (pełne w środku kółka).



## Pełny zestaw cech

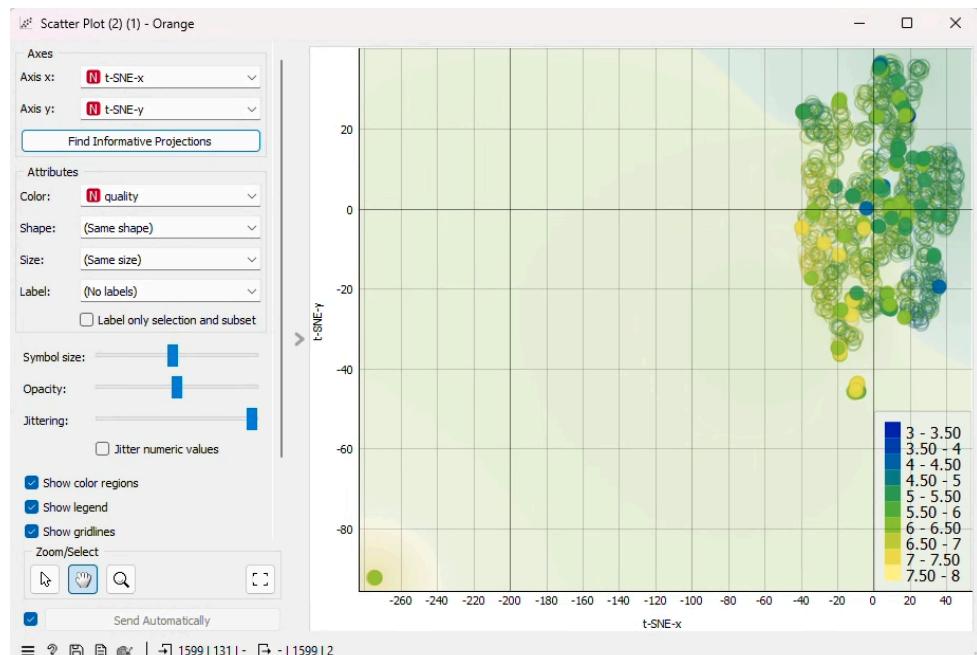
**Opis:** Wartości odstające (pełne kropki) są rozmieszczone głównie w dolnej i prawej części wykresu. Widać kilka punktów wyraźnie oddzielonych od głównych klastrów danych.

**Obserwacje:** W pełnym zestawie cech wartości odstające są bardziej rozproszone, co sugeruje, że różne cechy mogą wprowadzać szum, który utrudnia identyfikację skupisk.



## Podzbiór 3 najlepszych cech (przed odrzuceniem punktu skrajnego)

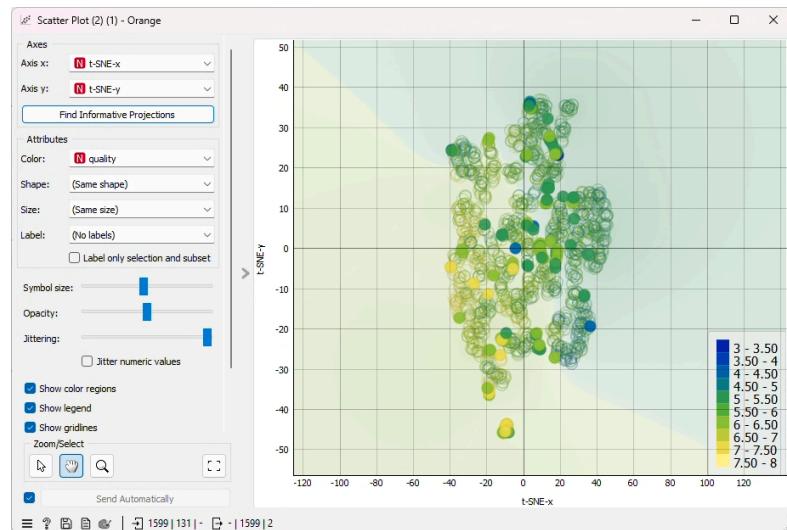
Widzimy, że punkt w lewym dolnym rogu jest wartością odstającą. W skupisku w prawym górnym rogu wykresu również występują outlery.



## Podzbiór 3 najlepszych cech (po odrzuceniu punktu skrajnego)

**Opis:** Wykres pokazuje bardziej zgrupowane punkty w centralnej części, z wyraźnie oddzielonymi punktami odstającymi (pełne kropki). Po odrzuceniu punktów z lewego dolnego narożnika, wizualizacja staje się bardziej klarowna.

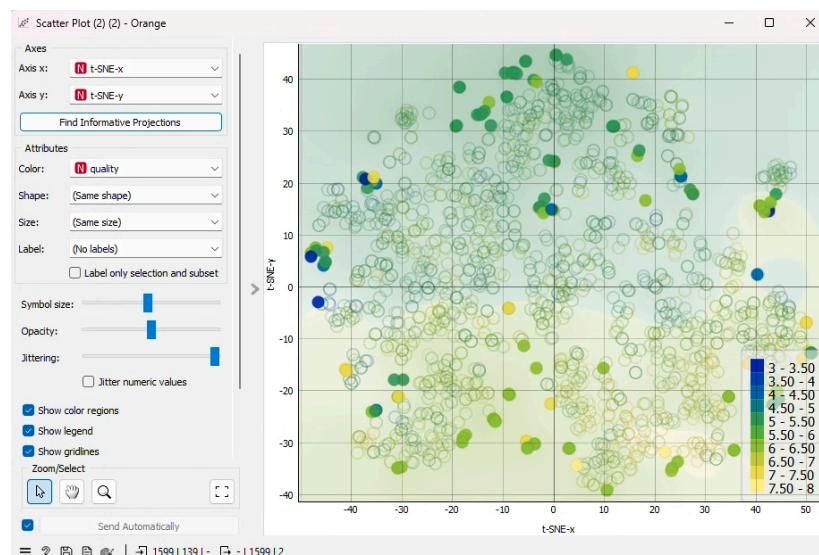
**Obserwacje:** Wyraźne punkty odstające są widoczne głównie w dolnej części wykresu. Usunięcie anomalii z lewego dolnego narożnika poprawiło czytelność wykresu i umożliwiło lepszą identyfikację klastrów.



## Trzy główne składowe (PC)

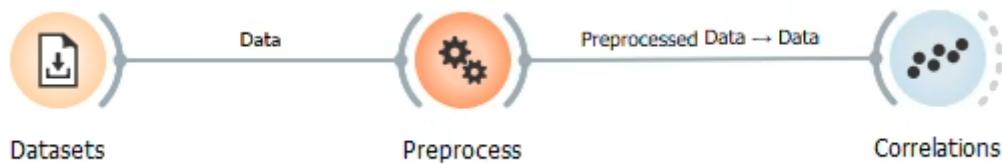
**Opis:** Wartości odstające są rozmiieszczane w różnych częściach wykresu, ale są mniej wyraźne niż w przypadku trzech najlepszych cech.

**Obserwacje:** Wykres dla głównych składowych PCA pokazuje bardziej równomierne rozmiesczenie punktów odstających, co może sugerować, że PCA skutecznie redukuje szum, ale może nie być tak skuteczne w identyfikacji wyraźnych anomalii.



## 4.5. Wizualizacja cech i sprawdzenie wniosków na temat ich podobieństwa

W celu zwizualizowania korelacji, wykorzystałem poniższy graf:



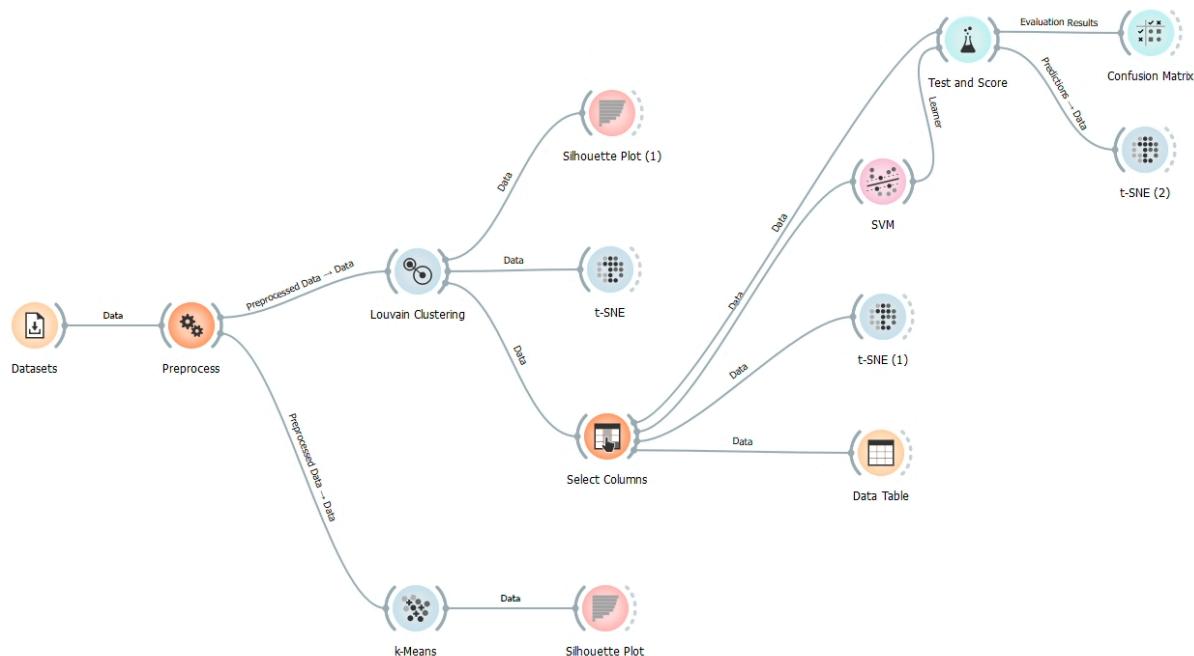
Przy pomocy widgetu Correlations, otrzymałem następujące korelacje posortowane od najsilniejszych do najsłabszych. Potwierdza to, że cechy takie jak **alcohol**, **volatile acidity** oraz **sulphates** są najsilniej skorelowane z jakością wina, co umacnia wcześniejsze rozważania.



## 4. Zadanie 4

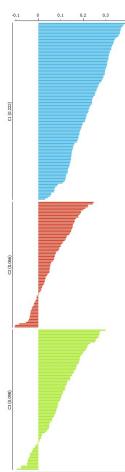
W tym zadaniu wykorzystałem zbiór "Heart Disease", ponieważ wcześniej używany zbiór "Wine quality - red" zawierał wartości numeryczne dla cechy **quality**, która jest cechą zależną od pozostałych, dlatego nie jest możliwa klasyfikacja dla tamtego zbioru. Aby wykonać to zadanie, wziąłem zbiór, dla którego mamy 2 klasy: chory, zdrowy.

Wykorzystałem poniższy graf do analizy:

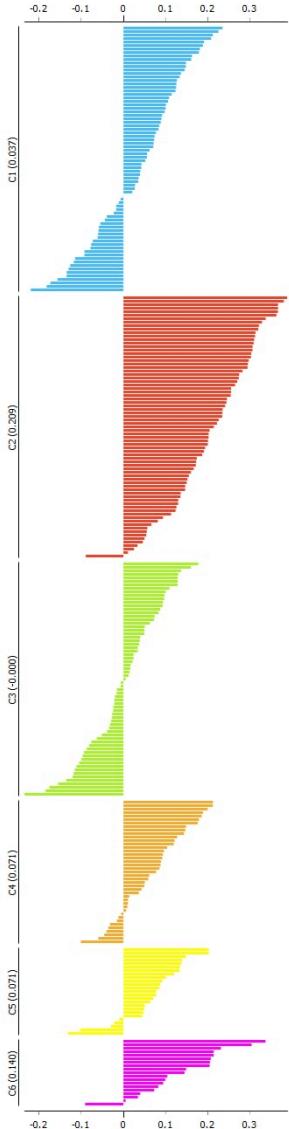


### 4.1. Wykresy Silhouette

**k-Means:** Wyniki wykresu Silhouette dla k-Means pokazują trzy klastry z wyższymi wartościami współczynnika Silhouette, co wskazuje na lepsze odseparowanie klastrów w porównaniu do Louvain Clustering.



**Louvain Clustering:** Wyniki wykresu Silhouette dla Louvain Clustering pokazują sześć klastrów z różnymi wartościami współczynnika Silhouette. Współczynniki dla większości klastrów są dodatnie, ale niektóre klastry mają wartości bliskie 0, co sugeruje, że punkty są na granicy klastrów.



## 4.2. Klasyfikacja

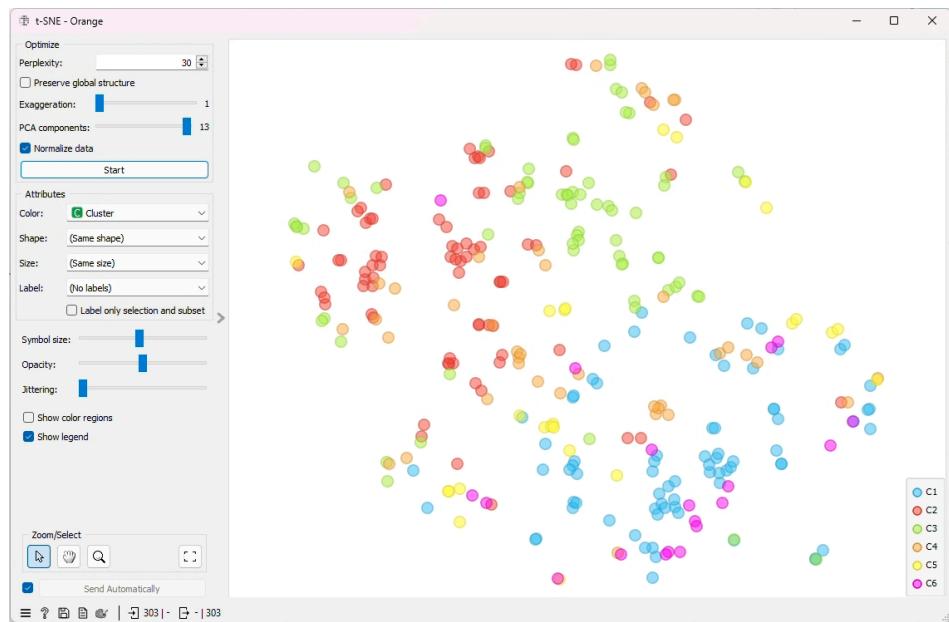
**SVM:** Użyłem klasyfikatora SVM do rekonstrukcji klastrów. Parametry SVM były ustawione zgodnie z poniższymi wartościami:

- **Kernel:** RBF (Radial Basis Function)
- **Cost (C):** 1.00
- **Gamma (g):** auto
- **Numerical tolerance:** 0.0010
- **Iteration limit:** 100

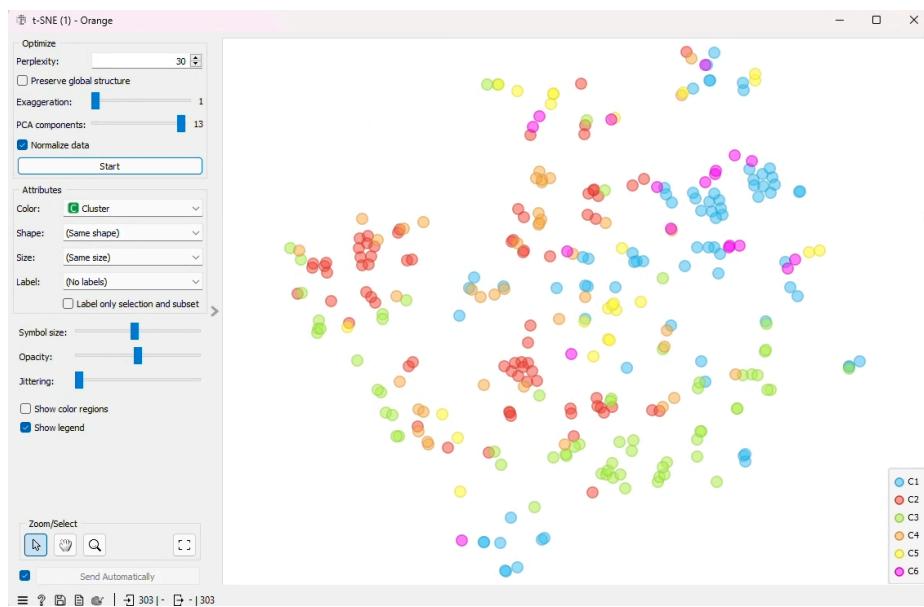
**Test and Score:** Przeprowadziłem ewaluację klasyfikatora za pomocą kroswalidacji (5-fold cross-validation).

## 4.3. Wizualizacja

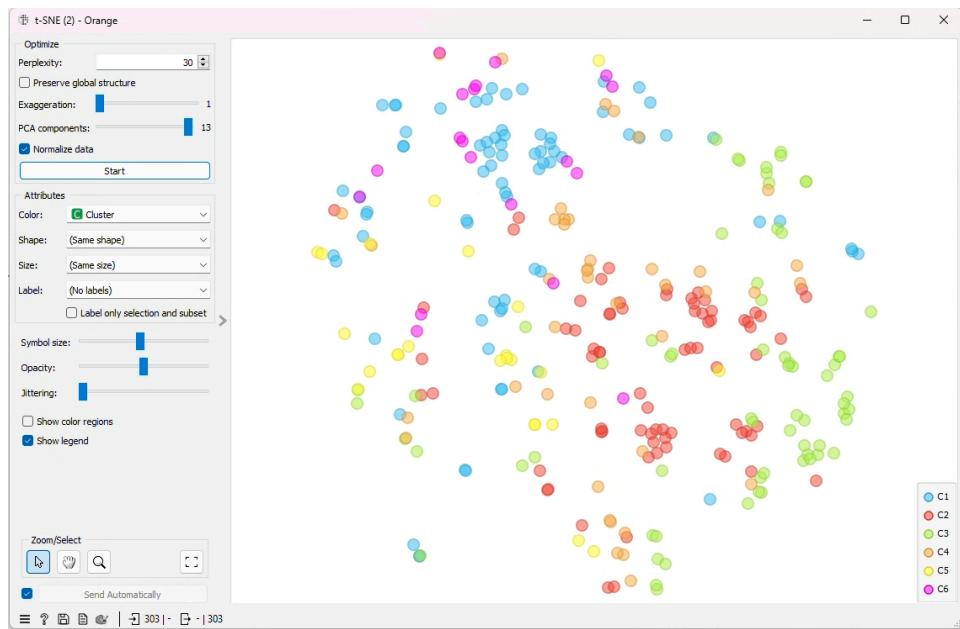
### t-SNE po Louvain Clustering



### t-SNE po Select Columns

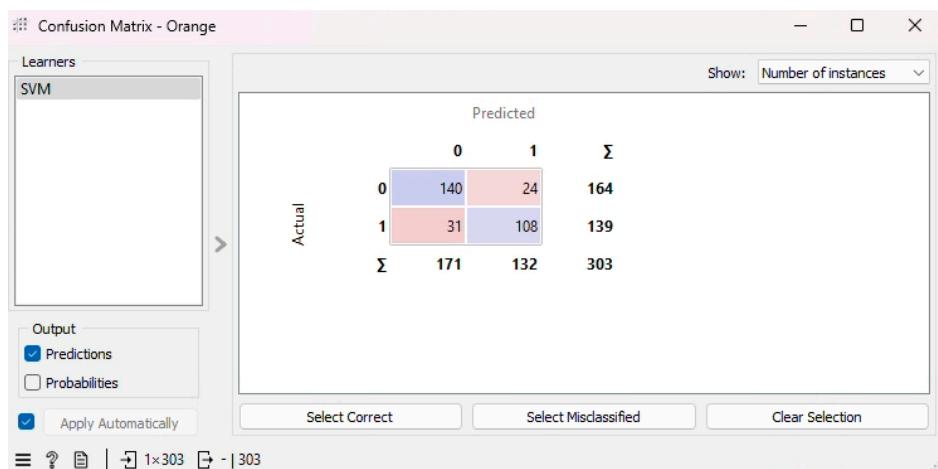


## t-SNE po Test and Score



## 4.4. Confusion Matrix

**Komentarz:** Macierz pomyłek pokazuje, że klasyfikator SVM skutecznie zrekonstruował klastry. Liczba błędów klasyfikacji jest niska, co potwierdza wysoką dokładność modelu.



## Wnioski:

1. SVM skutecznie zrekonstruował klastry i uzyskał wysokie wyniki w Test and Score oraz macierzy pomyłek, co czyni go odpowiednim klasyfikatorem do tego zadania.
2. t-SNE pozwoliło na wyraźne zobaczenie różnic między klastrami i klasami, potwierdzając skuteczność k-Means i SVM.