

Raport z Projektu: Statystyczna Analiza Danych

Czynniki chemiczne wpływające na jakość wina

Jakub Darul, Mateusz Lamla

21 stycznia 2026

Spis treści

1 Wstęp i cel analizy	3
2 Statystyki opisowe oraz korelacje	3
2.1 Podstawowe parametry statystyczne	3
2.2 Korelacje	4
3 Testy parametryczne (t-Studenta)	5
3.1 Sformułowanie hipotez	5
3.2 Wyniki	5
4 Analiza wariancji ANOVA	7
4.1 Sformułowanie hipotez	7
4.2 Wyniki	7
5 Testy nieparametryczne	9
5.1 Test Manna-Whitneya (Chlorki)	9
5.1.1 Sformułowanie hipotez	9
5.1.2 Wyniki	9
5.2 Test Kruskala-Wallisa (Kwasowość lotna)	10
5.2.1 Sformułowanie hipotez	10
5.2.2 Wyniki	10
6 Analiza regresji wielorakiej	11
6.1 Równanie modelu	11
6.2 Ocena modelu	11
7 Podsumowanie	12

1 Wstęp i cel analizy

Celem naszego projektu jest zbadanie, które z czynników chemicznych mają znaczący wpływ na oceny poszczególnych win. Analiza została przeprowadzona na zbiorze danych [WineQT.csv](#), składającym się z ponad tysiąca rekordów. Całość analizy skupiamy w głównej mierze na danych w kolumnie *Ocena*, które są subiektywnymi ocenami recenzentów w skali od 1 do 10. Każdy z testów, które przeprowadziliśmy, bada wpływ innej cechy chemicznej na oceny krytyków.

2 Statystyki opisowe oraz korelacje

Na początku wyznaczylismy podstawowe parametry statystyczne dla zmiennych numerycznych oraz macierz korelacji zmiennych.

2.1 Podstawowe parametry statystyczne

Skupimy się na cechach najistotniejszych oraz naszym zdaniem najciekawszych dla dalszej analizy:

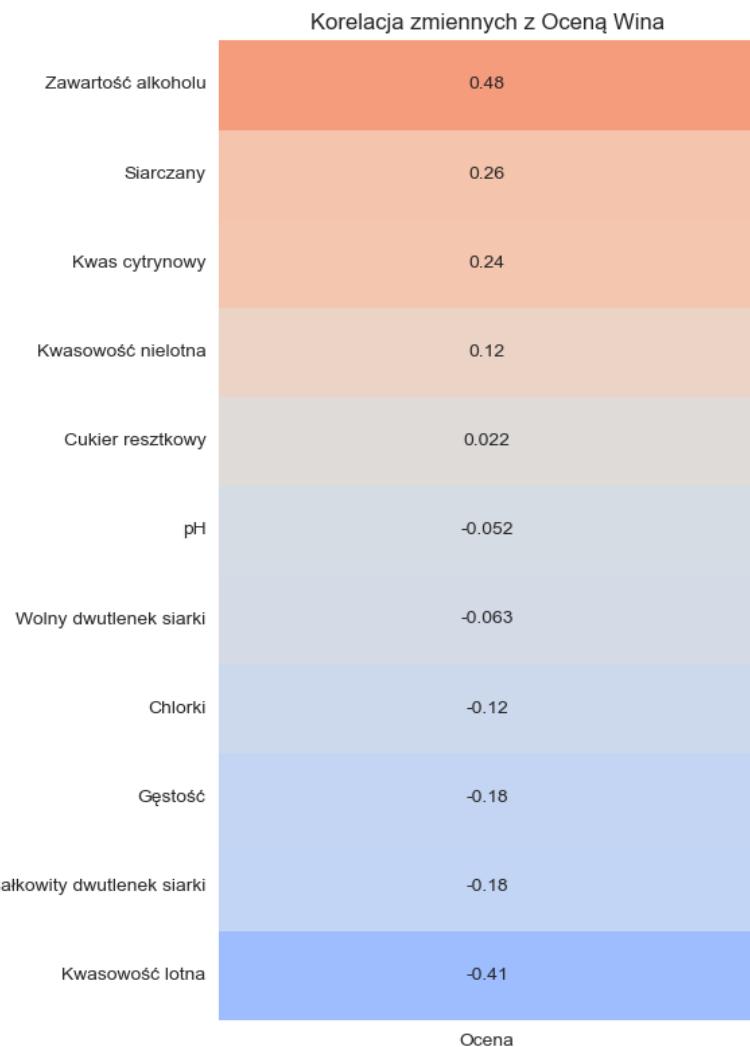
- **Zawartość alkoholu:** Średnia zawartość alkoholu dla win w zbiorze danych to 10.44% (z odchyleniem równym 1.08). Najmocniejsze wino w zbiorze ma w sobie aż 14.9% alkoholu, co znacząco odbiega od wartości średniej.
- **Ocena jakości:** Średnia ocena win dla całego zbioru to 5.66. Oceny są z zakresu od 3 do 8. Informacje te przydadzą się nam w późniejszej fazie projektu do kategoryzowania danych na potrzeby testów.

Reszta cech naszym zdaniem nie jest aż tak istotna, aby je opisywać dokładniej, dlatego zdecydowaliśmy się na jedynie umieszczenie wszystkich wartości w raporcie w formie tabeli.

	mean	std	min	max
Kwasowość nielotna	8.311111	1.747595	4.600000	15.900000
Kwasowość lotna	0.531339	0.179633	0.120000	1.580000
Kwas cytrynowy	0.268364	0.196686	0.000000	1.000000
Cukier resztkowy	2.532152	1.355917	0.900000	15.500000
Chlorki	0.086933	0.047267	0.012000	0.611000
Wolny dwutlenek siarki	15.615486	10.250486	1.000000	68.000000
Całkowity dwutlenek siarki	45.914698	32.782130	6.000000	289.000000
Gęstość	0.996730	0.001925	0.99007	1.00369
pH	3.311015	0.156664	2.74000	4.01000
Siarczany	0.657708	0.170399	0.33000	2.00000
Zawartość alkoholu	10.442111	1.082196	8.40000	14.90000
Ocena	5.657043	0.805824	3.00000	8.00000

Rysunek 1: Podstawowe parametry statystyczne zmiennych numerycznych

2.2 Korelacje



Rysunek 2: Korelacja zmiennych z oceną wina.

Z powyższej mapy korelacji (ograniczonej do korelacji z oceną) wynika, że najsilniejszą pozytywną korelację z oceną wina wykazuje **Zawartość alkoholu** (im wyższy procent alkoholu w winie, tym ocena wyższa). Najbardziej negatywna korelacja natomiast występuje dla **kwasowości lotnej** (im wyższa wartość, tym niższa ocena trunku). Ponadto, część z właściwości wina praktycznie wcale nie wpływa na ocenę końcową np. **cukier resztkowy** czy **pH**.

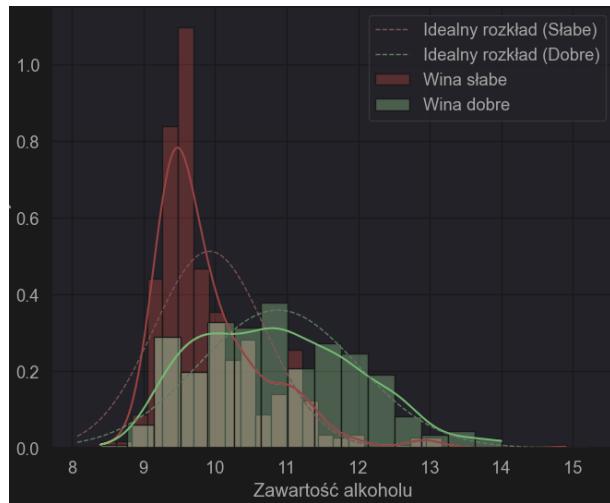
3 Testy parametryczne (t-Studenta)

3.1 Sformułowanie hipotez

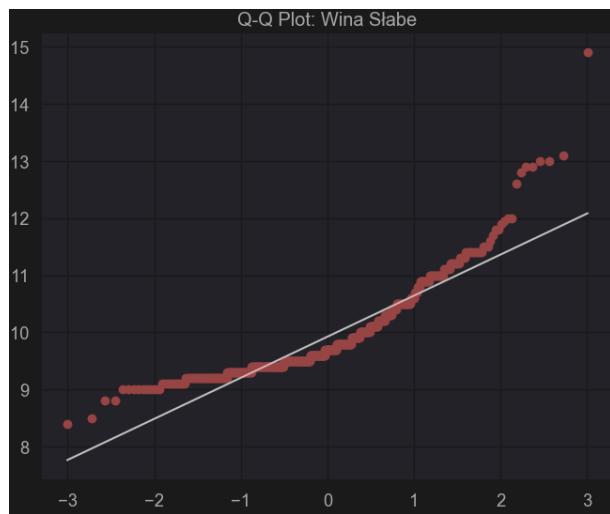
Pytanie badawcze: Czy wina sklasyfikowane jako ‘Dobre’ (ocena > 5) mają średnio wyższą zawartość alkoholu niż wina ‘Słabe’ (Ocena ≤ 5)?

- H_0 : Średnia zawartość alkoholu w obu grupach jest taka sama ($\mu_1 = \mu_2$).
- H_1 : Średnia zawartość alkoholu w winach ‘dobrych’ jest wyższa ($\mu_1 < \mu_2$).

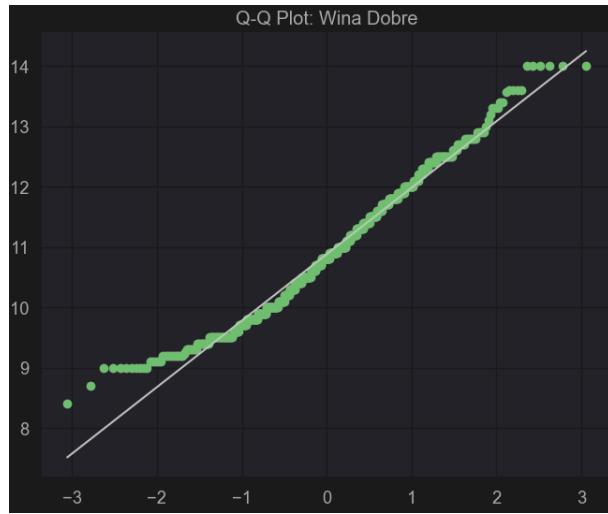
3.2 Wyniki



Rysunek 3: Histogram alkoholu



Rysunek 4: Q-Q wina słabe



Rysunek 5: Q-Q wina dobre

Przeprowadziliśmy test t-Studenta dla prób niezależnych, aby sprawdzić postawione wcześniej hipotezy. Test Shapiro-Wilka wykazał brak normalności rozkładu w obu grupach. Pomimo dużej liczebności prób i braku normalności, zastosowaliśmy test parametryczny powołując się na Centralne Twierdzenie Graniczne. Test Levene'a wykazał brak jednorodności wariancji, z powodu czego zastosowaliśmy test t-Studenta z poprawką Welch'a. Uzyskaliśmy poniższe wyniki:

- Średnia alkoholu (wina ‘Słabe’): 9.92
- Średnia alkoholu (wina ‘Dobre’): 10.88
- Statystyka t: -17.07
- Wartość p (p-value): ≈ 0.000

Wniosek: Ponieważ $p < 0.05$, hipoteza zerowa zostaje odrzucona. Różnica jest istotna statystycznie. Test wykazał, że wina oceniane wysoko mają wyższą zawartość alkoholu.

4 Analiza wariancji ANOVA

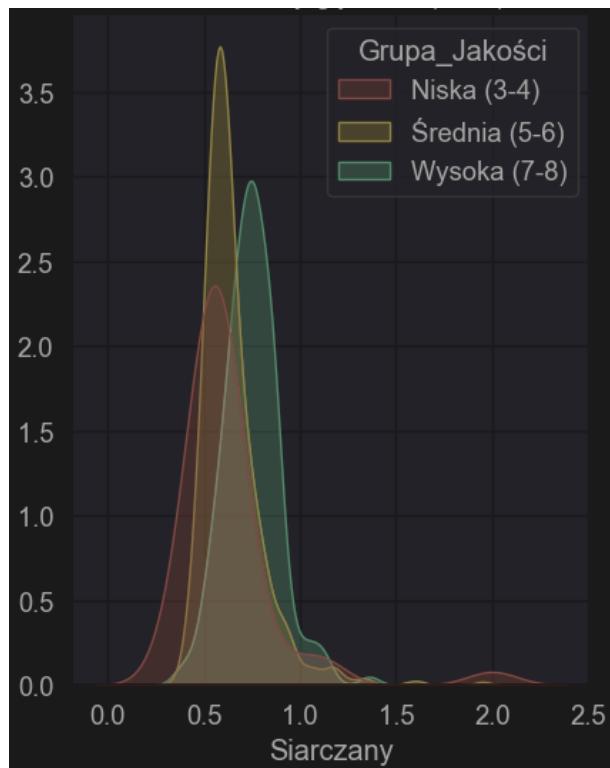
Zbadaliśmy wpływ grup jakościowych win na obecny w nich poziom siarczanów:

- Niska jakość (oceny 3-4)
- Średnia jakość (oceny 5-6)
- Wysoka jakość (oceny 7-8)

4.1 Sformułowanie hipotez

- H_0 : Średnie stężenie siarczanów jest równe we wszystkich grupach.
- H_1 : Przynajmniej jedna średnia różni się od pozostałych.

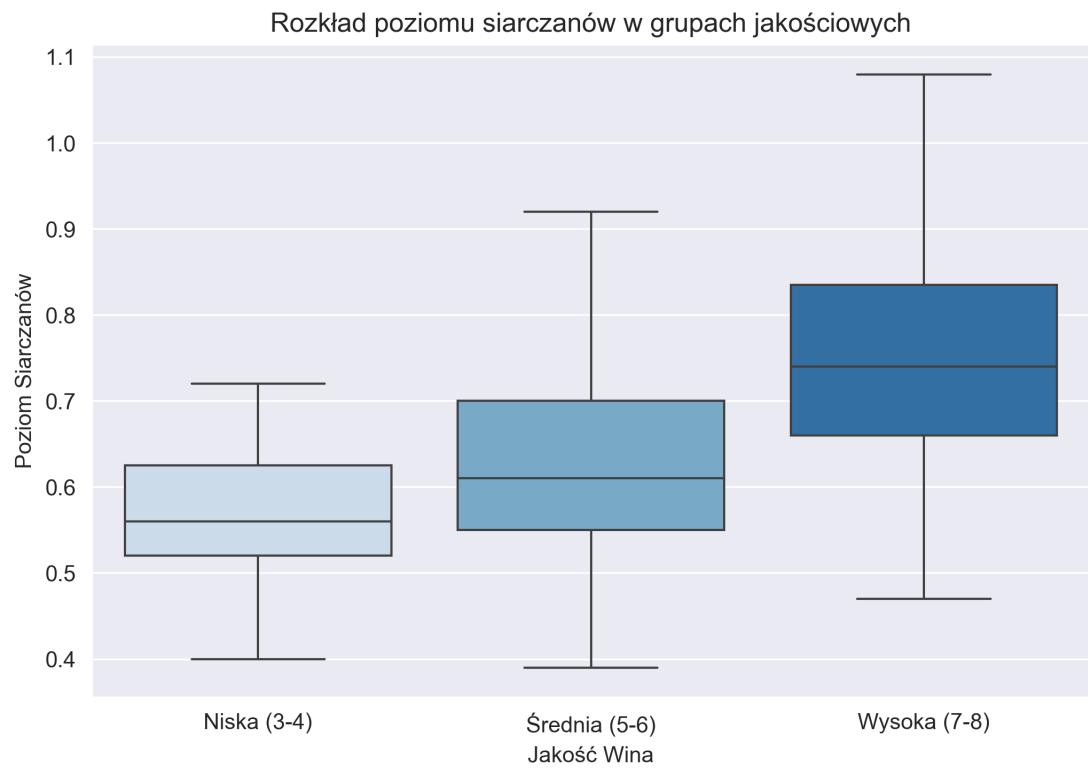
4.2 Wyniki



Rysunek 6: ANOVA gęstość

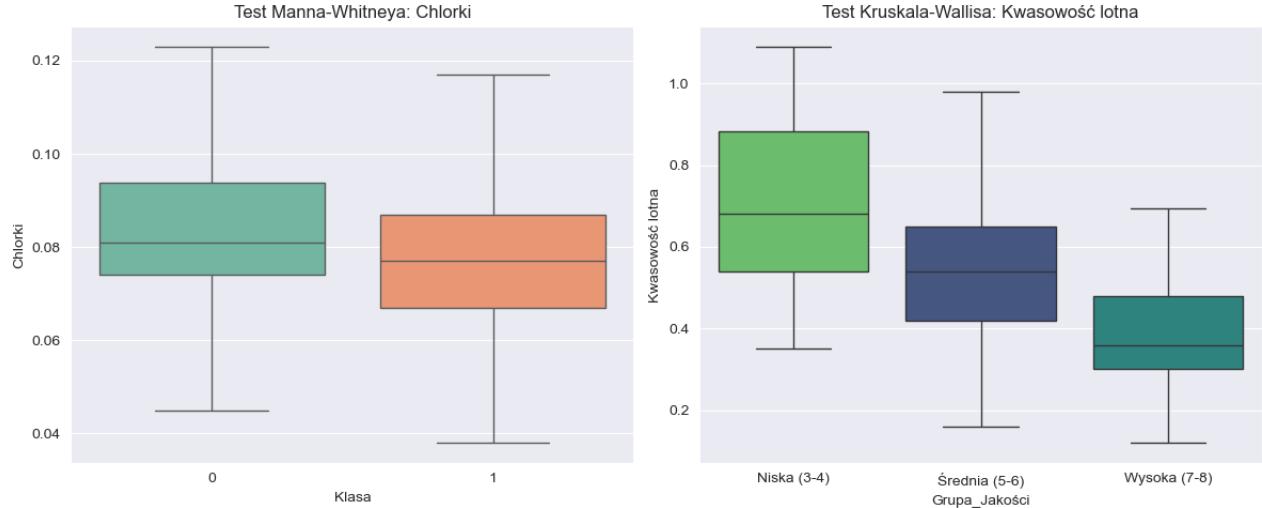
- Test Levene'a: ≈ 0.654
- Statystyka F: 26.06
- Wartość p (p-value): 8.58×10^{-12} ($p < 0.05$)

Wniosek: Test Levene'a wykazał, że założenie o jednorodności wariancji zostało spełnione ($p < 0.05$). Przeprowadzono test post-hoc Tukeya, który wykazał, że wina z wyższą oceną mają wyższy poziom siarczanów niż wina o średniej i niskiej ocenie. Między grupą średnią i niską nie odnotowano istotnych różnic. **Odrzucamy hipotezę zerową**, ponieważ występują istotne różnice między grupami.



Rysunek 7: Wykres ANOVA.

5 Testy nieparametryczne



Rysunek 8: Wykresy testów nieparametrycznych.

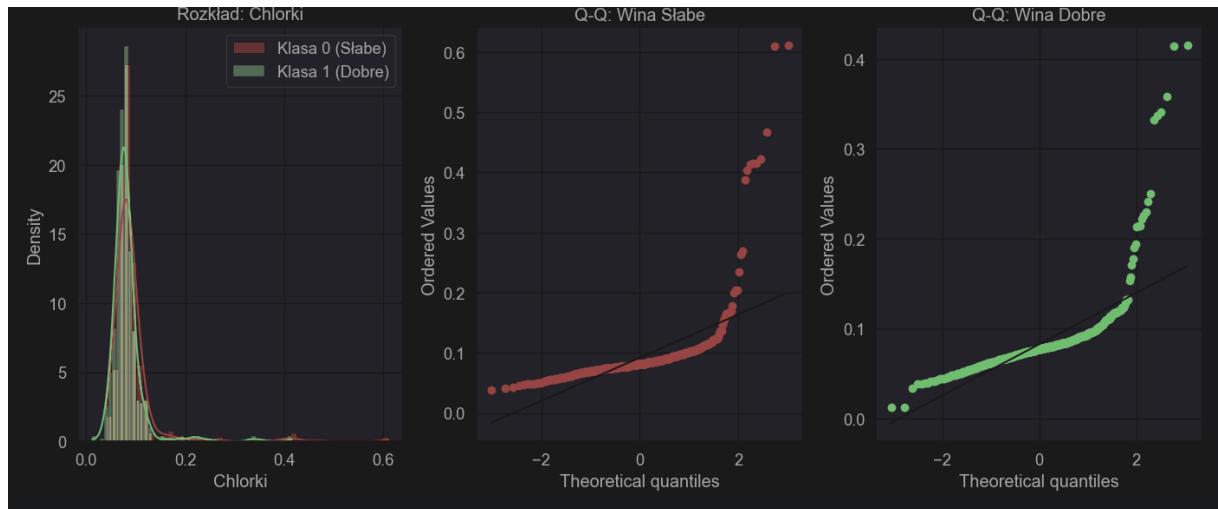
5.1 Test Manna-Whitneya (Chlorki)

Zbadaliśmy, czy rozkład poziomu chlorków różni się między winami ocenionymi jako gorsze, a winami ocenionymi jako lepsze.

5.1.1 Sformułowanie hipotez

- H_0 : Rozkłady poziomu chlorków w grupie win ‘słabych’ i win ‘dobrych’ są identyczne.
- H_1 : Rozkłady poziomu chlorków w obu grupach różnią się istotnie.

5.1.2 Wyniki



Rysunek 9: Wykresy testu Manna-Whitneya

- **p-value:** ≈ 0.000

Wniosek: Odrzucamy H_0 . Istnieje istotna statystycznie różnica w poziomie chlorków między grupami.

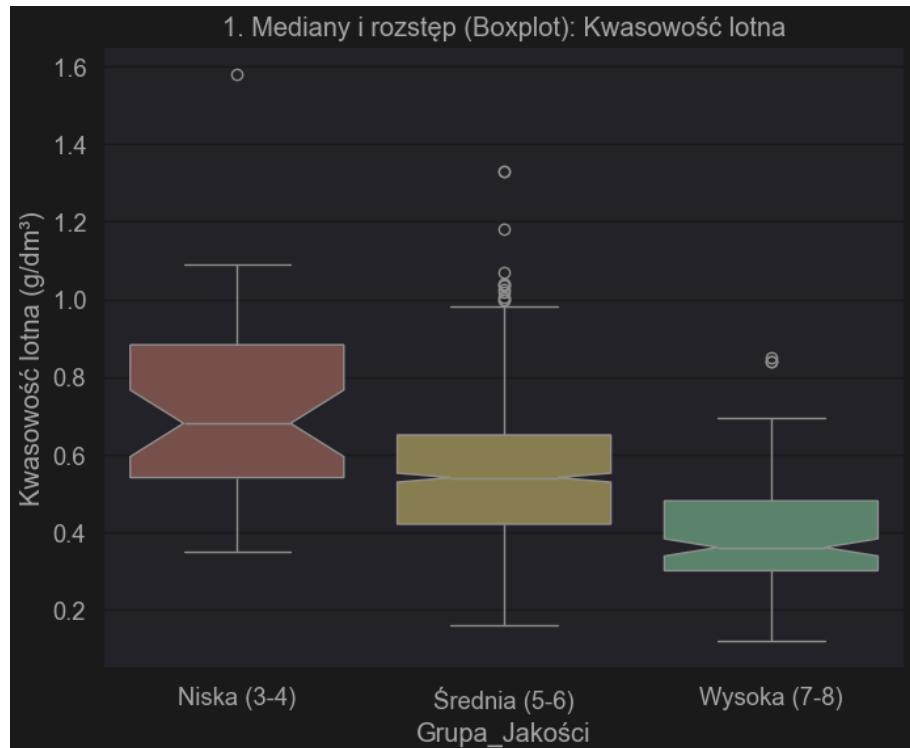
5.2 Test Kruskala-Wallisa (Kwasowość lotna)

Zbadaliśmy mediany kwasowości lotnej w trzech grupach jakościowych (słabe, średnie i dobre wina).

5.2.1 Sformułowanie hipotez

- H_0 : Mediany (rozkłady) kwasowości lotnej są równe we wszystkich analizowanych grupach jakościowych.
- H_1 : Przynajmniej jedna z grup różni się istotnie poziomem (medianą) kwasowości lotnej od pozostałych.

5.2.2 Wyniki



Rysunek 10: Mediany kwasowości lotnej

- **Statystyka H:** ≈ 142
- **p-value:** ≈ 0.000

Wniosek: Odrzucamy H_0 . Kwasowość lotna różni się istotnie pomiędzy grupami jakościowymi (jest niższa w winach ‘dobrej’ jakości).

6 Analiza regresji wielorakiej

Zbudowaliśmy model regresji liniowej w celu przewidzenia oceny wina na podstawie trzech zmiennych: **Zawartość alkoholu**, **Siarczany** oraz **Kwasowość lotna**.



Rysunek 11: Wykres regresji.

6.1 Równanie modelu

Na podstawie współczynników z regresji wielorakiej:

$$\text{Ocena} = 2.68 + 0.31 \cdot \text{Alkohol} + 0.66 \cdot \text{Siarczany} - 1.28 \cdot \text{Kwasowość Lotna} \quad (1)$$

6.2 Ocena modelu

- **R^2 (Współczynnik determinacji):** 0.353. Model wyjaśnia około 35.3% zmienności oceny wina.
- **Istotność zmiennych:** Wszystkie zmienne w modelu posiadają p-value < 0.05, co oznacza, że są istotne statystycznie.

7 Podsumowanie

Przeprowadzona analiza statystyczna wykazała, że końcowa ocena wina nie jest dziełem przypadku, a wynika z konkretnych parametrów chemicznych. Wina wysokiej jakości charakteryzują się wyższym poziomem zawartości alkoholu oraz siarczanów oraz znacznie niższym poziomem kwasowości lotnej i chlorków. Potwierdzeniem tego są zarówno testy parametryczne, jak i nieparametryczne, które wykonaliśmy. Utworzony przez nas model regresji pozwala oszacować, jaką ocenę uzyska wybrane przez nas wino, w oparciu o jego skład chemiczny. Wynika z niego, że kwasowość lotna jest głównym czynnikiem zdolnym do dyskwalifikacji wina jako produkt dobry dla konsumenta.