Analiza wybranych danych rzeczywistych z wykorzystaniem metod statystyki opisowej

Autorzy:

Marcelina Kosiorowska, Elżbieta Szopka

Politechnika
Wrocławska

Spis treści

1 Wstęp

Wiele państw na świecie posiada swoje unikalne alkohole. Nie inaczej jest w Portugalii, w której powstaje Vinho Verde. Jest to wyjątkowe wino znane z orzeźwiającego i owocowego aromatu. Zyskuje ono na popularności na całym świecie ze względu na swoje wyjątkowe walory smakowe. Nazwa „Vinho Verde" oznacza dosłownie „zielone wino" i odnosi się nie do koloru wina, bo może być ono białe, czerwone, a nawet różowe, ale do faktu, iż jest ono wytwarzane z niedojrzałych winogron. Dane niezbędne do przeprowadzenia wnikliwych obserwacji pozyskałyśmy za pomocą platformy Kaggle. Badano skład 1143 win w 12 kategoriach. Ze względu na obszerną ilość danych my skupimy się na 8 istotnych kategoriach. Są nimi: zawartość kwasów karboksylowych, kwas cytrynowy, cukier resztkowy, chlorki, gęstość wina, pH, zawartość alkoholu oraz jakość.

2 Podstawowe statystyki

2.1 Wzory oraz definicje:

2.1.1 Miary położenia

1. Średnia arytmetyczna, bardzo podatna na wartości odstające $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$,
2. Śrdnia harmoniczna $\bar{X}_{harm} = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$,
3. Średnia geometryczna $\bar{X}_{geo} = \sqrt[n]{x_1 x_2 \cdots x_n}$,
4. Średnia ucinana, czyli średnia arytmetyczna bez k najmnieszych i największych obserwacji $\bar{X}_{trim} = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} x_{(i)}$,
5. Średnia winsorowska, czyli średnia arytmetyczna ale k najmniejszych obserwacji jest zamieniana na obserwacje minimalną po obcięciu ich, analogicznie z wartościami największymi $\bar{X}_{wins} = \frac{1}{n}\left[(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}\right]$,
6. Mediana, czyli drugi kwartyl (Q2) $x_{med} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ jest nieparzyste} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & n \text{ jest parzyste} \end{cases}$,
7. Pierwszy kwartyl (Q1), czyli mediana obserwacji mniejszych od Q2,
8. Trzeci kwartyl (Q3), czyli mediana obserwacji większych od Q2.

2.1.2 Miary rozproszenia

1. Rozstęp międzykwartylowy, czyli różnica trzeciego kwartylu i pierwszego $IQR = Q3 - Q1$,
2. Wariancja z próby $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{X}\right)^2$,
3. Odchylenie standardowe z próby $S = \sqrt{S^2}$.

2.1.3 Pozostałe miary

1. Miara asymetrii, czyli współczynik skośności $\alpha = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{X}}{S} \right)^3$,

2. Miara spłaszczenia, inaczej kurtoza $K = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^4}{\left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{X})^2 \right)^2}$.

2.2 Tabele podstawowych statystyk dla danych wraz z interpretacją

Nazwa statystyki	Zaw. kw. karboksylowych	Kwas cytrynowy	pH	Chlorki
Średnia arytmetyczna	0.531	0.268	3.311	0.087
Średnia harmoniczna	0.469	0.104	3.304	0.077
Średnia geometryczna	0.501	0.21	3.307	0.081
Średnia ucinana	0.533	0.269	3.323	0.087
Średnia winsorowska	0.531	0.268	3.309	0.087
Wartość maksymalna	1.58	1	4.01	0.611
Wartość minimalna	0.12	0	2.74	0.012
Mediana	0.52	0.25	3.31	0.079
Pierwszy kwartyl (Q1)	0.3925	0.09	3.205	0.07
Trzeci kwartyl (Q3)	0.64	0.42	3.4	0.09
Rozstęp międzykwart. (IQR)	0.2475	0.33	0.195	0.02
Wariancja	0.032	0.039	0.025	0.0022
Odchylenie standardowe	0.18	0.197	0.157	0.047
Wsp. skośności (α)	0.681	0.371	0.221	6.018
Kurtoza	4.36	2.28	3.92	49.87

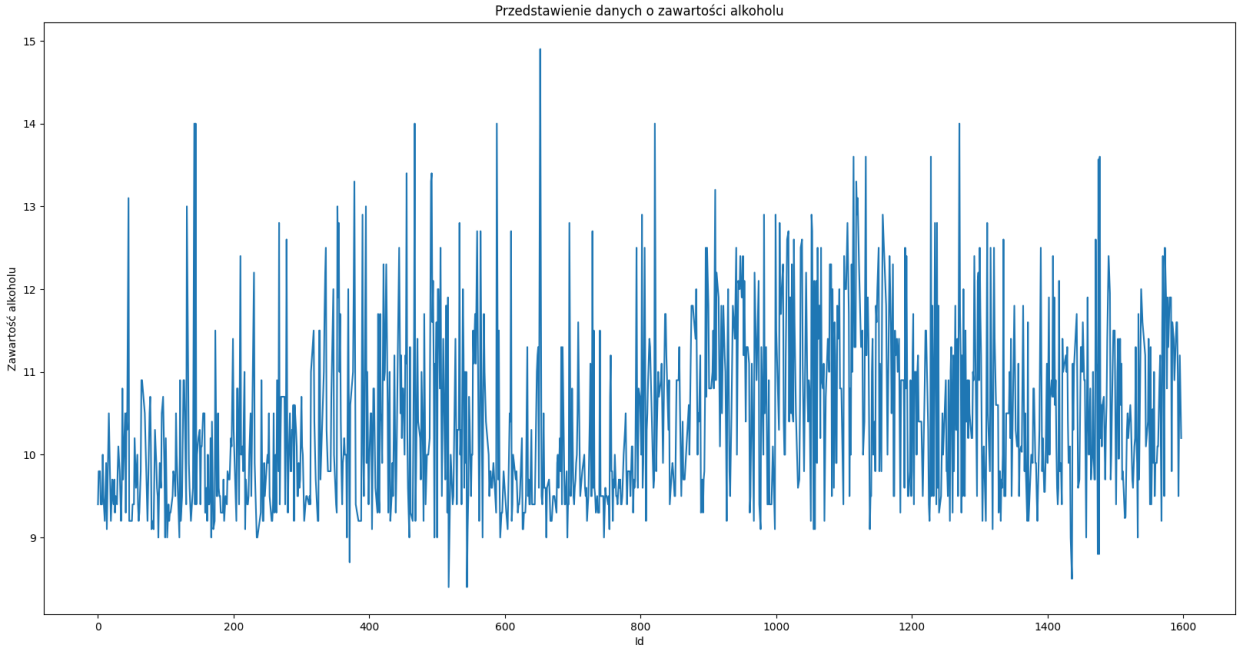
Nazwa statystyki	Cukier resztkowy	Gęstość wina	Zaw. alkoholu	Jakość
Średnia arytmetyczna	2.532	0.997	10.442	5.657
Średnia harmoniczna	2.221	0.997	10.338	5.541
Średnia geometryczna	2.340	0.997	10.389	5.6
Średnia ucinana	2.541	1.0002	10.479	5.677
Średnia winsorowska	2.533	0.996	10.434	5.654
Wartość maksymalna	15.5	1.00369	14.9	8
Wartość minimalna	0.9	0.99007	8.4	3
Mediana	2.2	0.997	10.2	6
Pierwszy kwartyl (Q1)	1.9	0.996	9.5	5
Trzeci kwartyl (Q3)	2.6	0.998	11.1	6
Rozstęp międzykwart. (IQR)	0.7	0.002	1.6	1
Wariancja	1.837	3.703	1.17	0.649
Odchylenie standardowe	1.355	0.002	1.082	0.805
Wsp. skośności (α)	4.355	0.102	0.862	0.286
Kurtoza	30.55	3.88	3.21	3.31

W powyższych tabelach zostały umieszczone wszystkie wykorzystywane przez nas statystyki dla każdej kategorii, w której badane było wino. Dane różnią się od siebie, jednak jesteśmy w stanie odczytać, że wykres każdej z badanych kategorii będzie prawostronnie skośny (dodatnie współczynniki skośności). Tak wysoka kurtoza dla chlorków oraz cukru resztkowego oznacza, że w danych jest więcej skrajnych wartości odstających niż w rozkładzie normalnym, co zobaczymy w dalszej części sprawozdania.

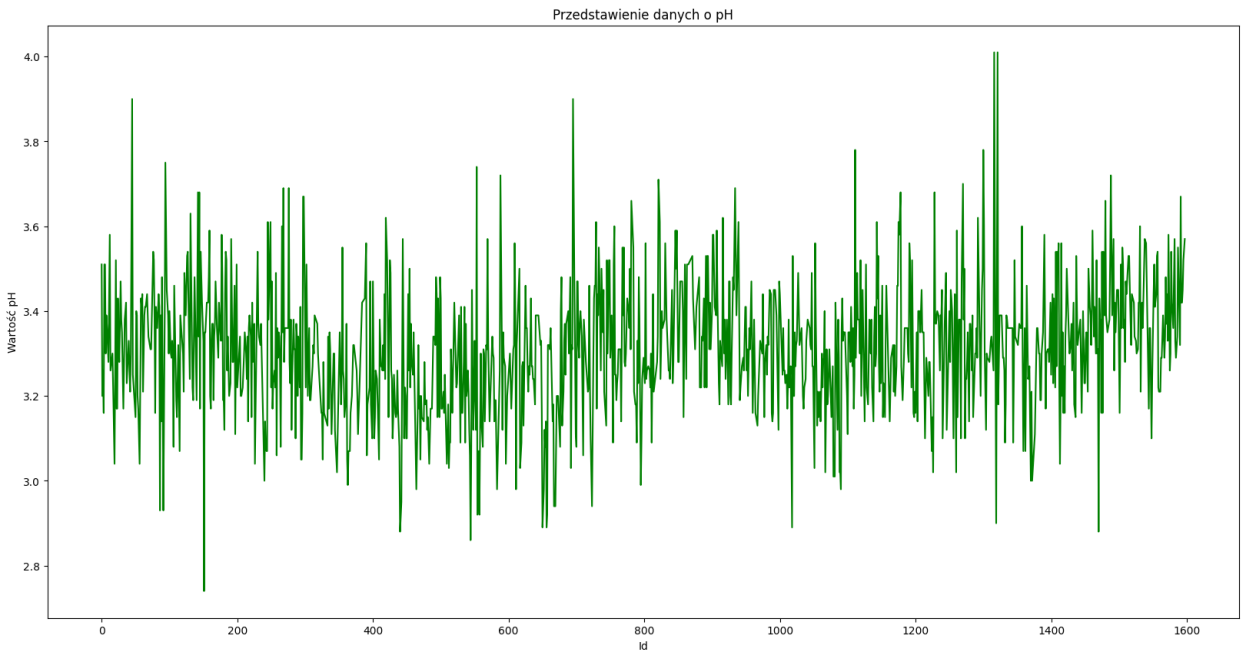
3 Wizualizacja danych

3.1 Przykładowe wykresy przedstawiające dane

Nasze rozważania na temat składu Vinho Verde rozpoczniemy od przedstawienia przykładowych danych. Dane na wykresach zostały przedstawione jedynie w kategoriach zawartości procentowej alkoholu oraz poziomu pH, ze względu na znaczną ilość danych. Przedstawione są one poniżej.



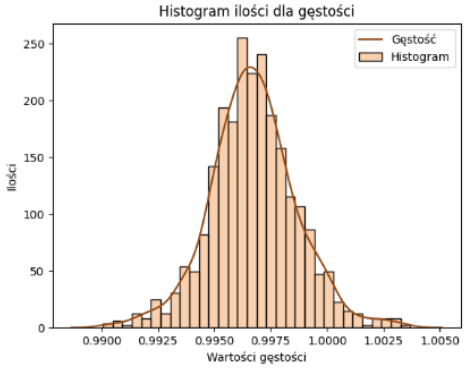
Wykres 1: Wykres zawartości procentowej alkoholu

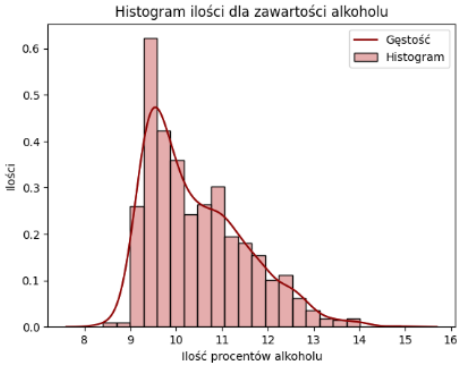


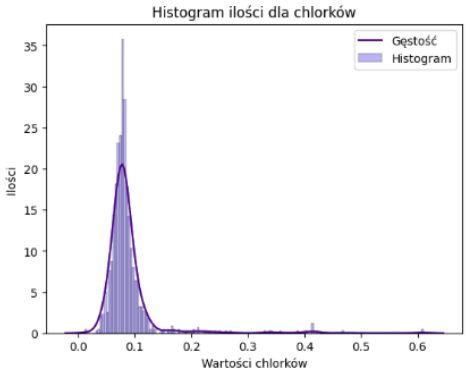
Wykres 2: Wykres poziomu pH w winie

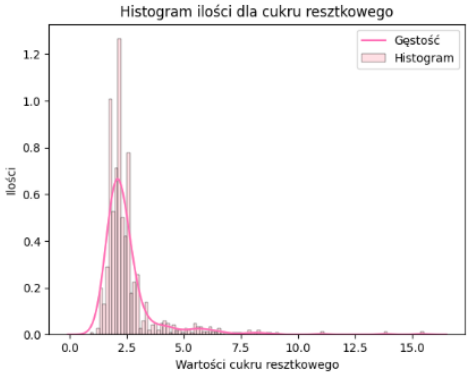
3.2 Histogramy oraz gęstości empiryczne

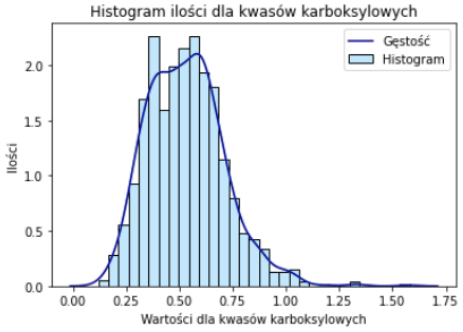
Teraz przejdziemy do przedstawienia histogramów oraz gęstości empirycznych dla naszych danych. Są one ciekawe ze względu na swój rozkład. Niektóre z nich wizualnie przypominają rozkład normalny, co w dalszej części zbadamy. Ilość badanych kategorii pokazuje i uświadamia nam, jak wiele czynników składa się na produkcję wina, tak jak i każdego innego produktu. Jest to niezwykle złożony proces, na szczęście dzięki statystykom jesteśmy w stanie go lepiej zrozumieć.

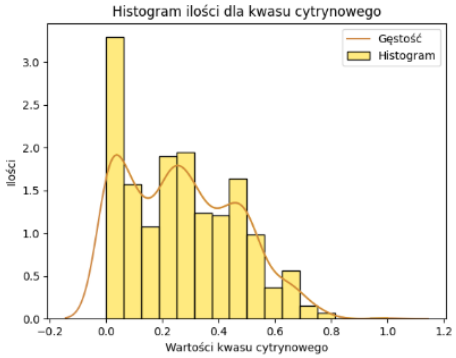


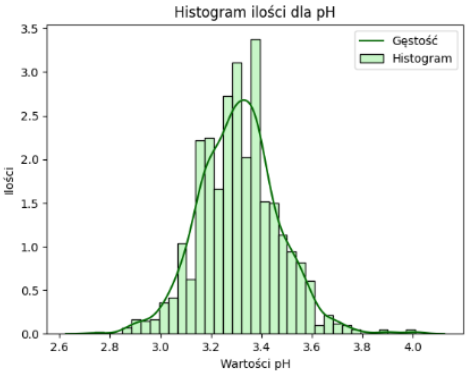


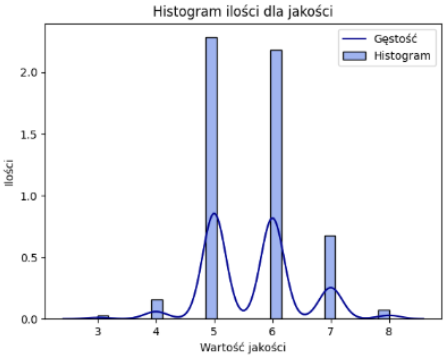








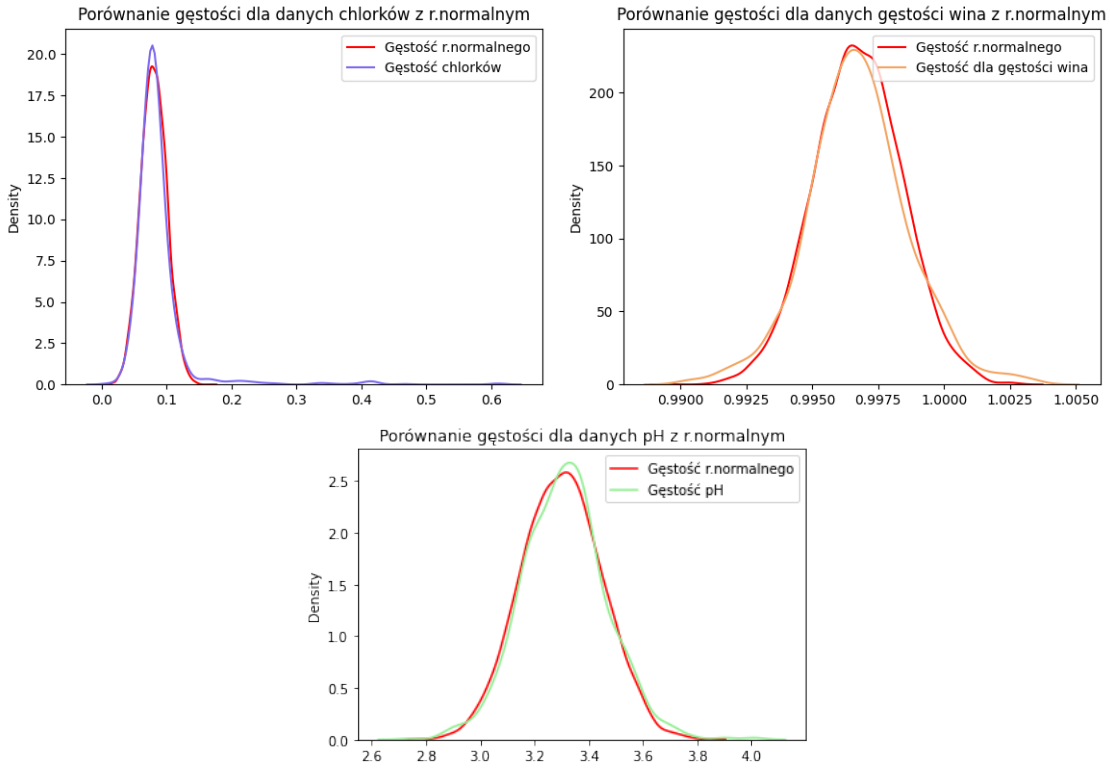




Z powyższych wykresów jesteśmy w stanie wywnioskować, że gęstość wina jest zbliżona do gęstości wody (tj. $1\frac{g}{cm^3}$). Widzimy również, że zawartość procentowa alkoholu w tym unikalnym dla Portugalii trunku wynosi najczęściej około 9,5 procent, co jest dość popularną wartością. Jeśli chodzi o cukier resztkowy to przedstawimy krótką definicję. Cukier resztkowy to taki cukier, który pozostaje w winie po zakończeniu procesu fermentacji. Poziom takiego cukru określamy w gramach na litr ($\frac{g}{l}$). Przyjmuje się, że gdy cukier resztkowy występuje w ilości do $4\frac{g}{l}$ to wino możemy nazywać winem wytrawnym, stąd też wniosek, że Vinho Verde jest wytrawne. Możemy również zauważyć, że w tym alkoholu występują chlorki oraz kwas cytrynowy. Chlorki wspomagają działanie kwasów, stąd też niskie, czyli kwasowe, pH wina. Każdy ze zbadanych czynników ma wpływ na jakość. W tym badaniu ustalono konkretne wartości dla jakości, czyli od 3 do 8. Zauważmy, że zdecydowana większość próbek mieści się w wartościach 5 i 6, czyli są to głównie wina o poprawnej, średniej jakości.

3.3 Porównanie z rozkładem normalnym

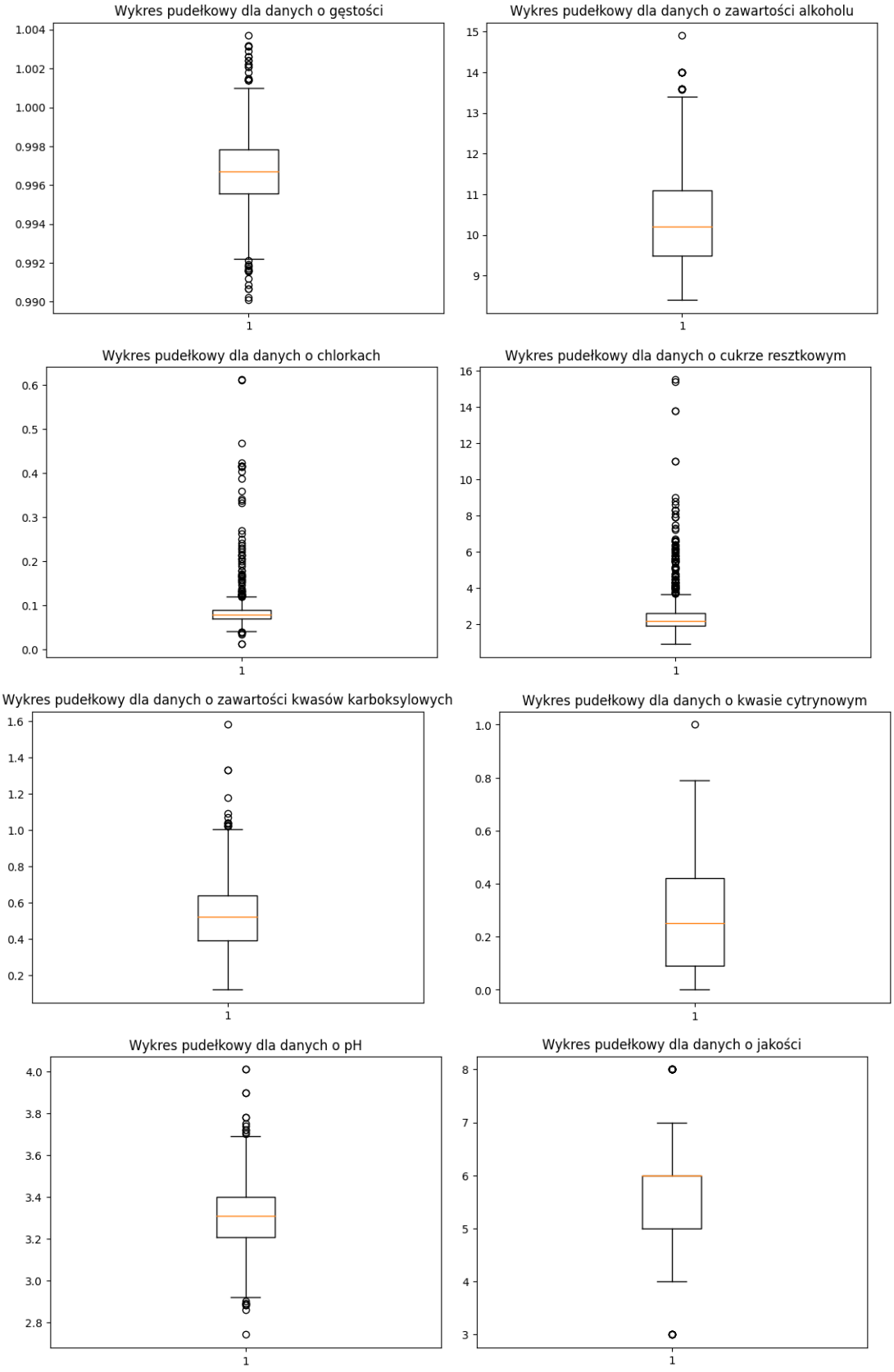
Gęstości empiryczne niektórych danych wizualnie przypominały nam gęstość rozkładu normalnego. Są to gęstości dla: pH, gęstości wina oraz chlorków. Zobaczmy jak to porównanie wygląda na wykresach.



Postanowiłyśmy zbadać ten fakt przy pomocy testu Shapiro-Wilka. W Pythonie funkcja scipy.stats.shapiro() pozwala na wykonanie owego testu. Funkcja ta zwraca wynik testu oraz wartość p, która jest miarą istotności statystycznej.
Poziom istotności ustalono na 0,05. Wyniki testu Shapiro-Wilka obejmują statystykę testu i wartość p. Wartość p można porównać z poziomem istotności, aby określić, czy można odrzucić hipotezę o normalności rozkładu. Jeśli wartość p jest większa niż poziom istotności, to nie ma wystarczających dowodów na odrzucenie hipotezy o normalności rozkładu. W przeciwnym przypadku hipoteza o normalności rozkładu zostaje odrzucona.
U nas okazało się, że każdy badany przypadek został odrzucony. Zatem rozkłady te są jedynie zbliżone wizualnie do rozkładu normalnego.

3.4 Boxploty

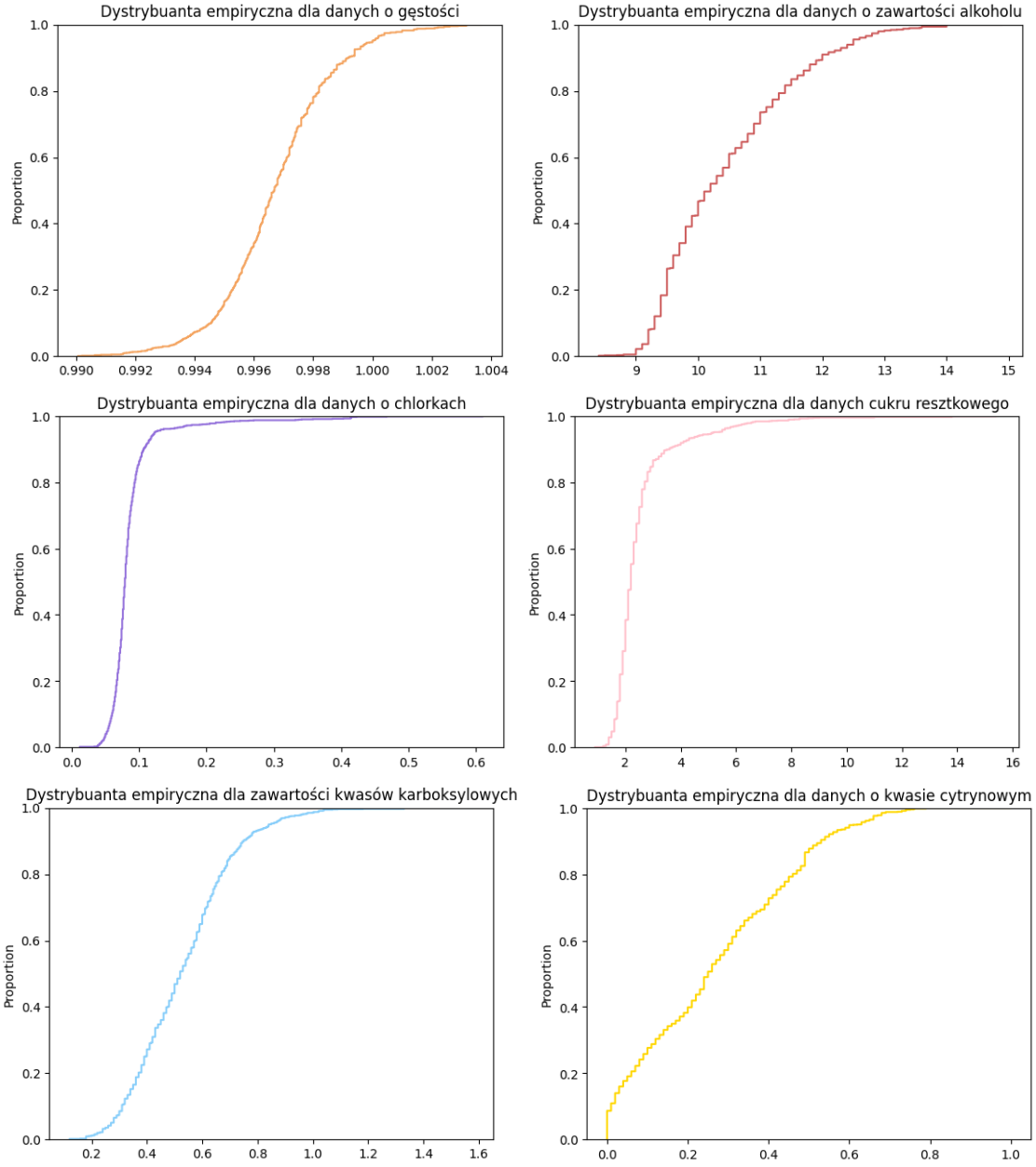
Teraz przeanalizujmy jak nasze dane zachowują się przedstawione na wykresach pudełkowych.

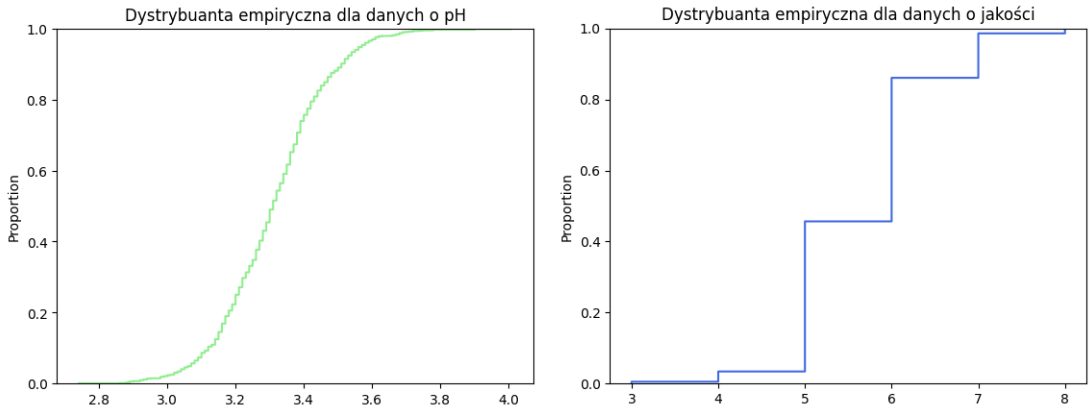


Analizując powyższe dane możemy zauważyć, że największy rozstęp kwartylowy występuje w danych o zawartości alkoholu, a najmniejszy w danych o gęstości wina, co zgadza się z obliczeniami i wynikami przedstawionymi w tabelach. Najliczniejsze obserwacje odstające występują na wykresach danych o chlorkach oraz cukrze resztkowym. To zgadza się z wcześniejszymi obliczeniami kurtozy. Wykres pudełkowy danych o gęstości jest równie symetryczny, jak histogram tej kategorii.

3.5 Dystrybuanty

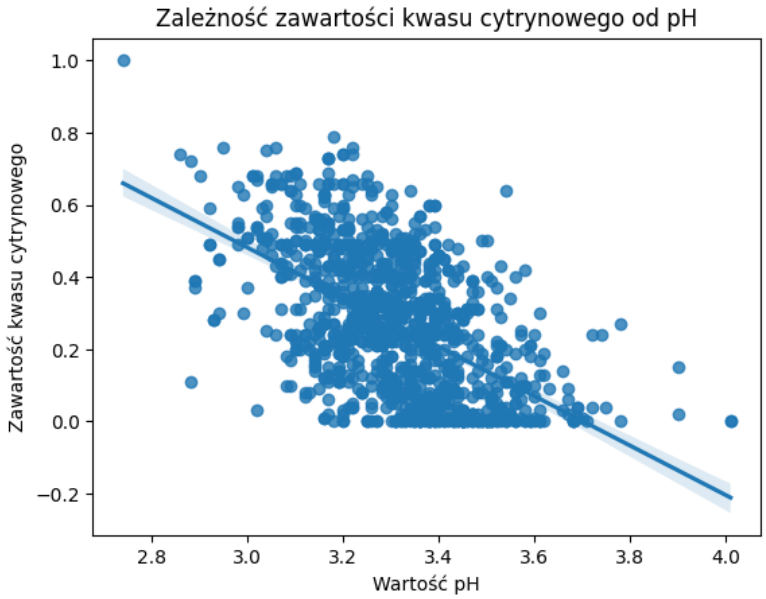
Poniżej przedstawimy również dystrybuanty dla każdej z grup danych, co dopełni informacje na ich temat.



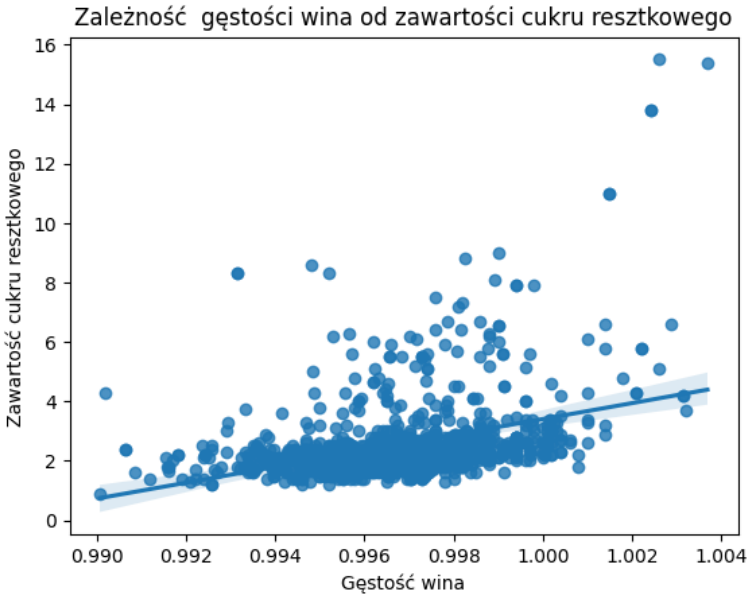


4 Powiązanie danych ze sobą

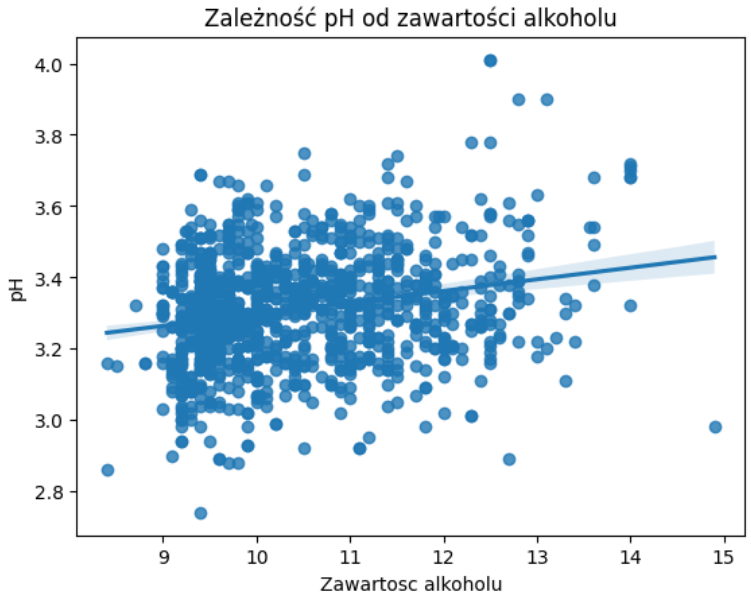
Po przeanalizowaniu wszystkich wykresów oraz statystyk pojawia się pytanie - co to wszystko tak naprawdę oznacza?
Proces fermentacji i produkcji produktu związany jest z wieloma czynnikami. Im więcej czynników w danym procesie występuje, tym bardziej będą one ze sobą powiązane. Zainteresowane tematem postanowiłyśmy stworzyć kilka wykresów punktowych wraz z liniami trendu, które pozwoliły nam lepiej zrozumieć omawiany temat. Poniżej przedstawimy kilka z nich.



Rozpoczniemy od najbardziej intuicyjnej zależności. Skala pH przyjmuje wartości od 0 do 14, gdzie poziomy od 0 do 6 informują nas o kwasowym charakterze roztworu. Stąd też oczywisty wniosek, że im mniej kwasu cytrynowego, tym wyższe pH, ponieważ zmniejsza się ilość jonów kwasowych (H+).



Gęstość wina ma związek z ilością cukru resztkowego, ale ta zależność nie jest jednoznaczna. Cukier resztkowy to cukier, który pozostaje w winie po zakończonej fermentacji, czyli po przemianie cukrów na alkohol przez drożdże. Wina o wyższej zawartości cukru resztkowego zwykle mają wyższą gęstość niż wina o niższej zawartości cukru resztkowego, ponieważ cukier ma większą masę cząsteczkową niż alkohol i inne składniki wina.



Ostatnia zależność jest najmniej jednoznaczna, ponieważ zależy ściśle od danych, a nie od zachodzących procesów chemicznych. Może to wynikać z wielu czynników, takich jak jakość i poziom kwasowości wina, odmiana winogron, warunki klimatyczne podczas uprawy, czas zbioru i wiele innych. Zwiększenie zawartości alkoholu w winie może prowadzić do zmniejszenia kwasowości i zwiększenia pH, ale jeśli inne czynniki w danym przypadku przeważają, to pH może wzrosnąć wraz ze wzrostem zawartości alkoholu.

5 Podsumowanie

Badanym przez nas zestawem danych był skład portugalskiego Vinha Verde. Podzieliłyśmy dane na 8 kategorii. Efekty wykonanych obliczeń zaprezentowałyśmy pod postacią wartości liczbowych w przejrzystych tabelach. Ponadto zwizualizowałyśmy dane przy pomocy różnego rodzaju wykresów. Zdołałyśmy w ten sposób wyciągnąć liczne wnioski oraz pokazać zależności między kategoriami danych. Dowiedziałyśmy się też ja skomplikowanym procesem jest fermentacja wina. Ta dogłębna analiza przeprowadzona na danych Vinho Verde jest również dowodem jak potężnym narzędziem są elementy statystyki opisowej.

6 Bibliografia

1. Źródło danych (Wine Quality Dataset - Vinho Verde)