Vícerozměrná analýza dat

GitHub Repo

Aplikace PCA, shlukové analýzy a LDA na chemických vlastnostech vín

Autor: Michal Šeda
Dataset: Wine Recognition Data (UCI ML Repository)
Metody: PCA, K-means, Hierarchické shlukování, LDA

Úvod a formulace úlohy

Popis úlohy

Cílem tohoto projektu je aplikovat metody vícerozměrné analýzy dat na reálný dataset chemických vlastností vín z různých odrůd. Dataset obsahuje **178 vzorků vín** ze tří různých odrůd révy vinné pěstovaných ve stejné oblasti Itálie. Pro každý vzorek je k dispozici **13 chemických proměnných**.

Zdroj dat

UCI Machine Learning Repository - Wine Recognition Data

Dataset: 178 vzorků vín ze 3 odrůd révy vinné z Itálie
Proměnné: 13 chemických parametrů

Přehled parametrů Wine datasetu

1. ALCOHOL (Alkohol)

- **Jednotka:** % obj.
- **Popis:** Obsah ethanolu ve víně. Vyšší hodnoty = silnější víno. Ovlivňuje chuť, tělo a stabilitu.
- **Interpretace:** ↑ = silnější víno, vyšší extraktivita
- **Typický rozsah:** 11-15%

2. MALIC ACID (Kyselina jablečná)

- **Jednotka:** g/L
- **Popis:** Hlavní organická kyselina v hroznech. Ovlivňuje kyselost a čerstvost vína. Klesá při dozrávání hroznů.
- **Interpretace:** ↑ = kyselejší, čerstvější víno
- **Poznámka:** Během malolaktické fermentace se přeměňuje na kyselinu mléčnou

3. ASH (Popel)

- **Jednotka:** g/L
- **Popis:** Anorganické minerály zbylé po spálení vína. Indikátor minerálního složení půdy.
- **Interpretace:** ↑ = vyšší minerální obsah

- **Obsahuje:** K, Ca, Mg, Na, Fe a další minerály

4. ALCALINITY OF ASH (Alkalinita popela)

- **Jednotka:** meq/L (miliekvivalenty na litr)
- **Popis:** Schopnost popela neutralizovat kyseliny. Souvisí s obsahem draslíku, vápníku a hořčíku.
- **Interpretace:** ↑ = vyšší pH, nižší kyselost
- **Souvisí s:** Pufrační kapacitou vína

5. MAGNESIUM (Hořčík)

- **Jednotka:** mg/L
- **Popis:** Důležitý minerál z půdy. Ovlivňuje enzymatické reakce při fermentaci. Indikátor terroir.
- **Interpretace:** ↑ = bohatší půda, lepší terroir
- **Rozsah:** 70-162 mg/L v datasetu

6. TOTAL PHENOLS (Celkové fenoly)

- **Jednotka:** g/L
- **Popis:** Suma všech fenolických sloučenin. Antioxidanty, ovlivňují barvu, chuť a stárnutí vína.
- **Interpretace:** ↑ = větší antioxidační kapacita
- **Zahrnuje:** Flavonoidy + neflavonoidní fenoly
- **Význam:** Nejvyšší u červených vín

7. FLAVANOIDS (Flavonoidy)

- **Jednotka:** g/L
- **Popis:** Podskupina fenolů. Hlavní antioxidanty, ovlivňují barvu (červená), chuť (trpkost) a zdravotní přínosy.
- **Interpretace:** ↑ = silnější barva, více tříslovin
- **Zahrnuje:** Katechiny, anthokyaniny, quercetin
- **KLÍČOVÝ PARAMETR pro rozlišení odrůd!**

8. NONFLAVANOID PHENOLS (Neflavonoidní fenoly)

- **Jednotka:** g/L
- **Popis:** Fenoly bez flavonoidní struktury. Menší vliv na barvu, ale přispívají k celkové chuti.
- **Interpretace:** ↓ = čistší flavonoidní profil
- **Zahrnuje:** Kyselina gallová, kyselina kávová, tyrosol

9. PROANTHOCYANINS (Proanthokyaniny)

- **Jednotka:** mg/L
- **Popis:** Typ flavonoidů (kondenzované taniny). Zodpovědné za trpkost a "tělnatost" červených vín. Stabilizují barvu.
- **Interpretace:** ↑ = trpčí, tělnatější víno
- **Pocit:** "Sucho" v ústech po červeném víně

10. COLOR INTENSITY (Intenzita barvy)

- **Jednotka:** bezrozměrná (absorbance)
- **Popis:** Měří intenzitu/sytost barvy vína. Vyšší u červených vín, nižší u bílých.
- **Interpretace:** ↑ = tmavší, sytější barva

- **Měření:** Spektrofotometricky

11. HUE (Odstín)

- **Jednotka:** bezrozměrná (poměr absorbancí)
- **Popis:** Odstín/tón barvy.
  - Nízké hodnoty = červená/fialová (mladá vína)
  - Vysoké hodnoty = oranžová/hnědá (oxidace, stárnutí)
- **Interpretace:** ↓ = mladší, ↑ = starší/oxidované
- **Výpočet:** OD420/OD520

12. OD280/OD315 OF DILUTED WINES (OD280/OD315 zředěných vín)

- **Jednotka:** poměr (bezrozměrný)
- **Popis:** Poměr absorbancí UV světla při 280 nm a 315 nm. Indikátor obsahu proteinů a fenolů.
- **Interpretace:** ↑ = vyšší kvalita, více proteinů a fenolů
- **Proč zředěné?** Čisté víno je příliš tmavé pro přesné měření
- **OD280:** Zachycuje proteiny + fenoly
- **OD315:** Zachycuje flavonoidy + barevné látky
- **RYCHLÝ TEST kvality vína!**

13. PROLINE (Prolin)

- **Jednotka:** mg/L
- **Popis:** Aminokyselina. Nejvíce zastoupená aminokyselina ve víně (až 85%). Ovlivňuje nutriční hodnotu a chuť.
- **Interpretace:** ↑ = nutriční hodnota, plnost
- **Rozsah:** 278-1680 mg/L v datasetu (obrovská variabilita!)
- **Zajímavost:** Některé odrůdy mají 10× vyšší obsah než jiné

Skupiny parametrů

1. ZÁKLADNÍ SLOŽENÍ

- **Alcohol** - hlavní alkohol
- **Malic acid** - hlavní kyselina
- **Ash, Alcalinity, Magnesium** - minerální složení (terroir)

2. FENOLICKÉ SLOUČENINY

(Antioxidanty, barva, chuť)

- **Total phenols** - celkový obsah fenolů
- **Flavanoids** - hlavní barevné a třískovinové látky
- **Nonflavanoid phenols** - doplňkové fenoly
- **Proanthocyanins** - třísloviny (trpkost)

3. OPTICKÉ VLASTNOSTI

- **Color intensity** - síla barvy
- **Hue** - odstín barvy (mladé vs. staré)
- **OD280/OD315** - spektrofotometrický ukazatel kvality

4. AMINOKYSELINY

- **Proline** - hlavní aminokyselina

Praktické využití

Pro vinaře:

- **Flavonoidy + Total phenols** → kvalita, potenciál stárnutí
- **Malic acid** → čerstvost, vhodnost pro malolaktickou fermentaci
- **Proline** → typické pro určité odrůdy
- **Hue** → monitorování oxidace/stárnutí

Pro analytiky:

- **OD280/OD315** → rychlý test kvality (nemusí se měřit všechny fenoly samostatně)
- **Magnesium + Alkalinita** → "otisk prstu" půdy/terroir
- **Color intensity + Hue** → vizuální kvalita

Pro spotřebitele:

- **Alcohol** → síla vína
- **Proanthocyanins** → trpkost červených vín
- **Total phenols** → antioxidační přínosy

Reference Jackson, R.S. (2008). Wine Science: Principles and Applications

Explorační analýza dat (EDA)

Metadata datasetu	
Vlastnost	**Hodnota**
Počet vzorků	178
Počet proměnných	13

Seznam sledovaných odrůd
Odrůda 1
Odrůda 2
Odrůda 3

Rozdělení vzorků v třídách	
Odrůda	**Počet vzorků**
Odrůda 1	59
Odrůda 2	71
Odrůda 3	48

Popisná statistika

h	magnesium	total_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_inte
0	178.00	178.00	178.00	178.00	178.00	1
9	99.74	2.30	2.03	0.36	1.59	
4	14.28	0.63	1.00	0.12	0.57	
0	70.00	0.98	0.34	0.13	0.41	
0	88.00	1.74	1.21	0.27	1.25	
0	98.00	2.35	2.13	0.34	1.56	
0	107.00	2.80	2.88	0.44	1.95	

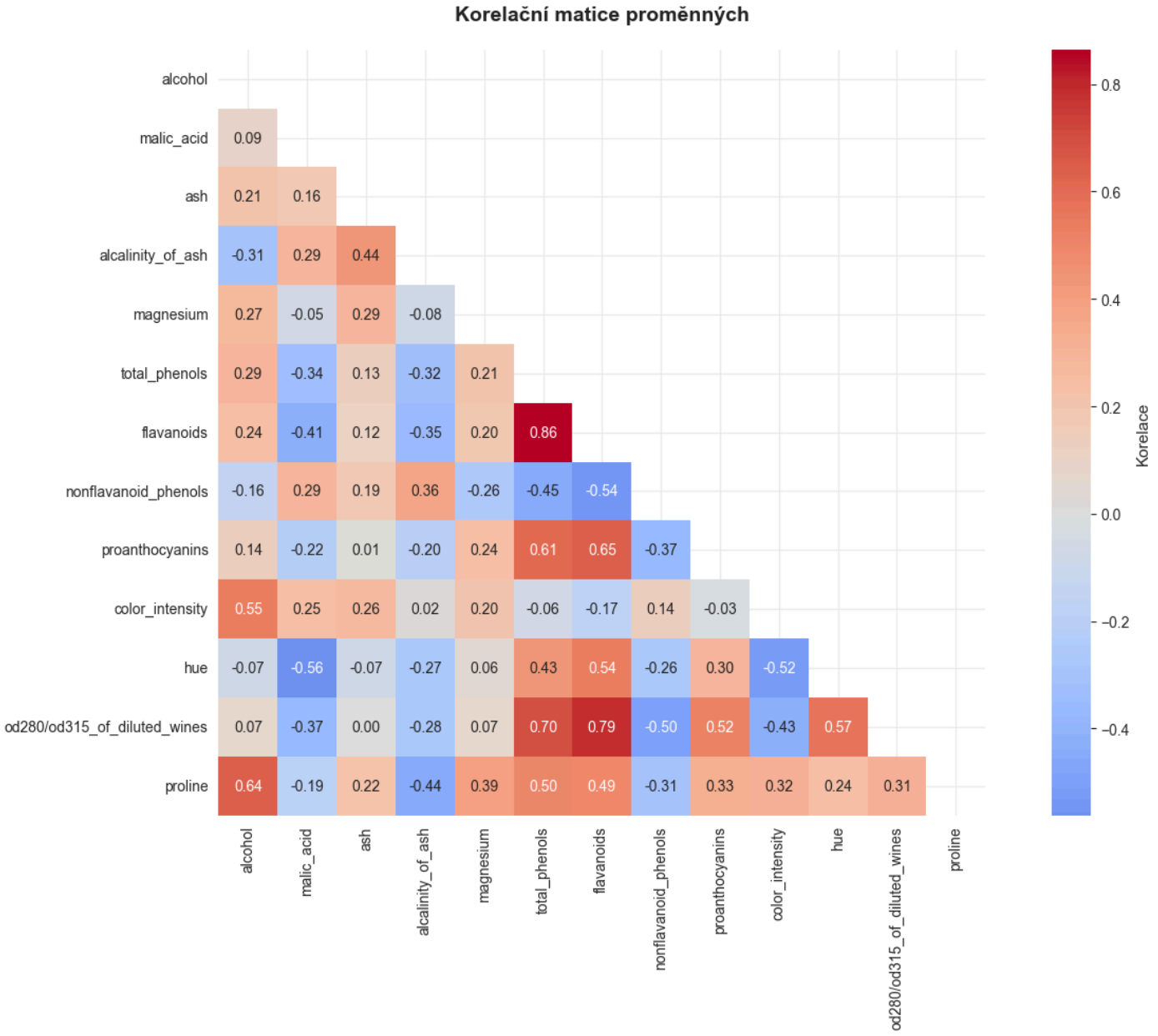
Popisná statistika

Ukázka dat

ɔtal_phenols	flavanoids	nonflavanoid_phenols	proanthocyanins	color_intensity	hue	od28
2.8	3.06	0.28	2.29	5.64	1.04	
2.65	2.76	0.26	1.28	4.38	1.05	
2.8	3.24	0.3	2.81	5.68	1.03	
3.85	3.49	0.24	2.18	7.8	0.86	
2.8	2.69	0.39	1.82	4.32	1.04	

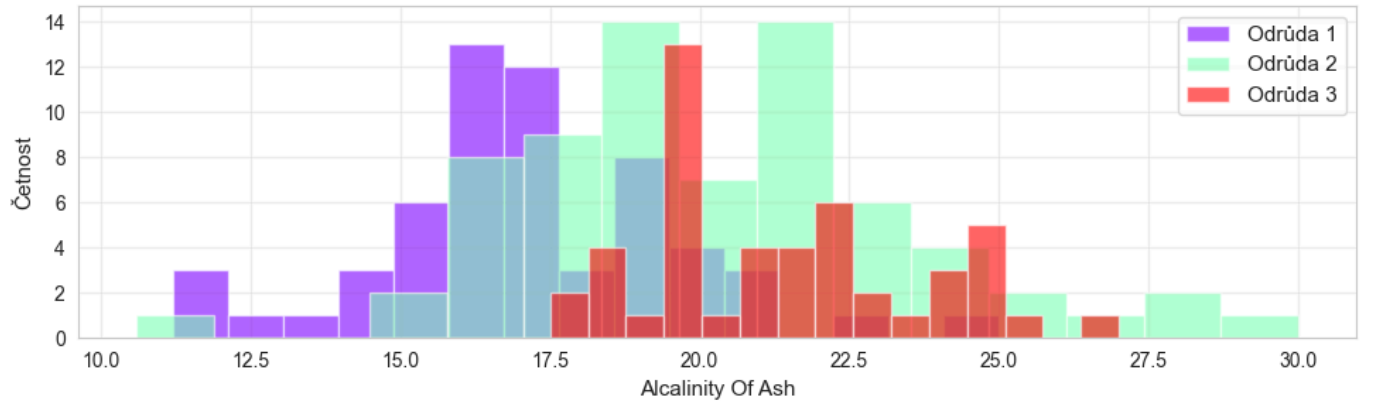
Chybějící hodnoty: 0

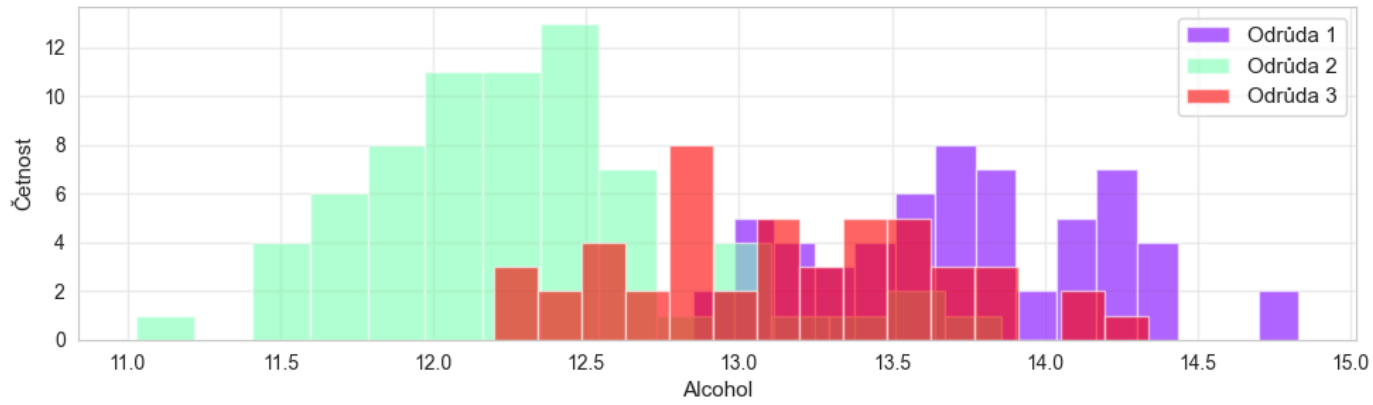
Dataset neobsahuje chybějící hodnoty

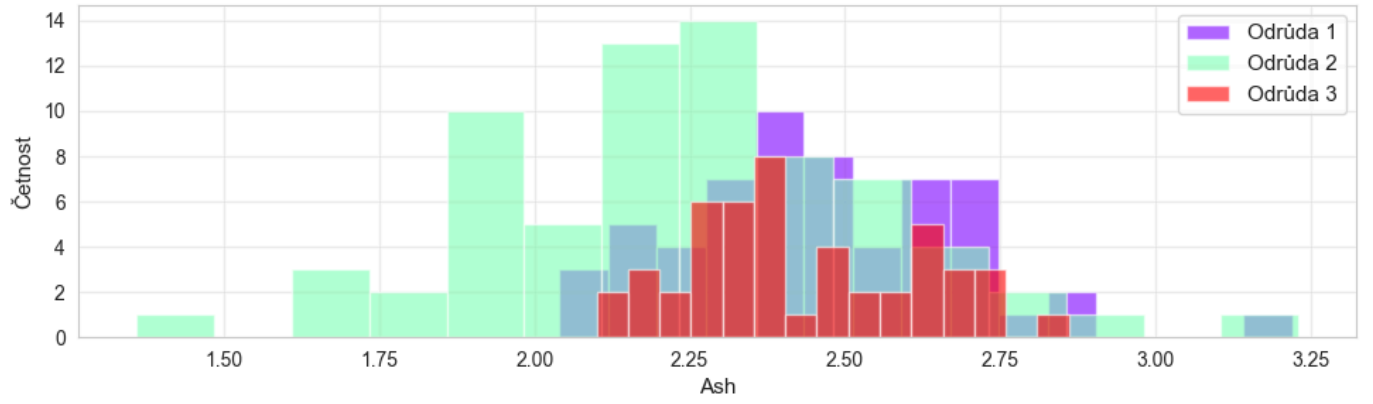


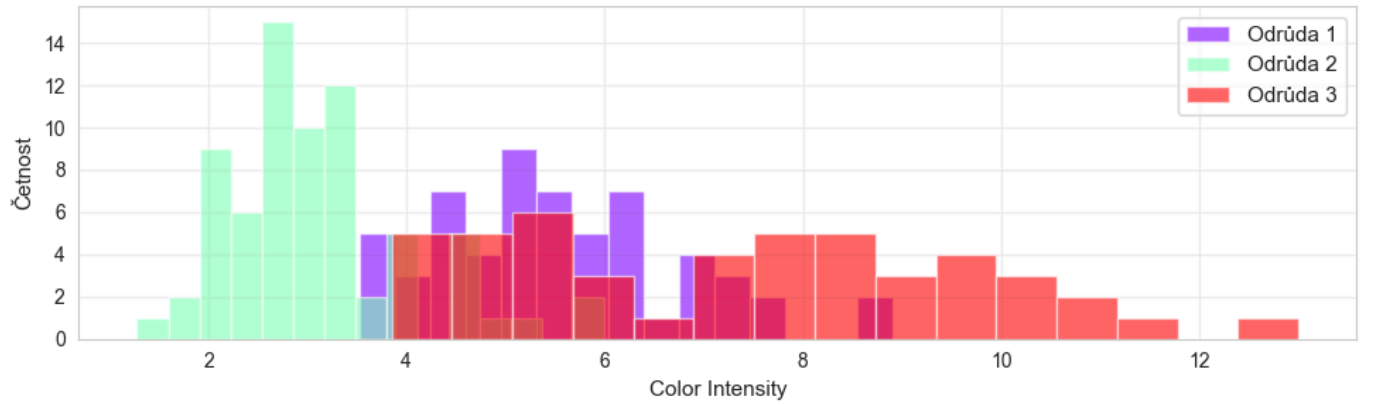
Nejsilnější pozitivní korelace		
		Korelace
total_phenols	flavanoids	0.8646
flavanoids	od280/od315_of_diluted_wines	0.7872
total_phenols	od280/od315_of_diluted_wines	0.6999
flavanoids	proanthocyanins	0.6527
alcohol	proline	0.6437

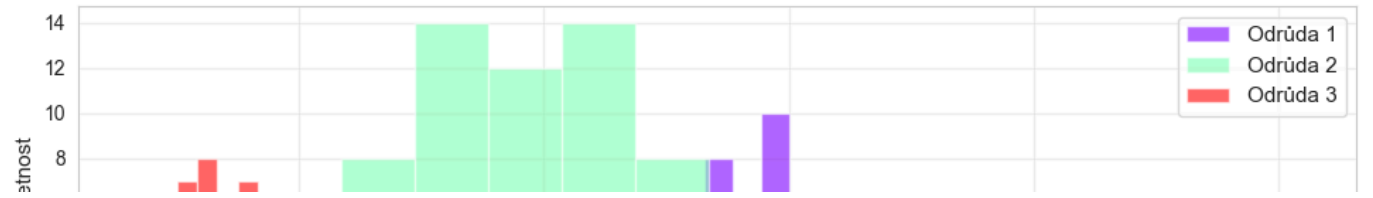
Rozdělení vybraných proměnných podle odrůd

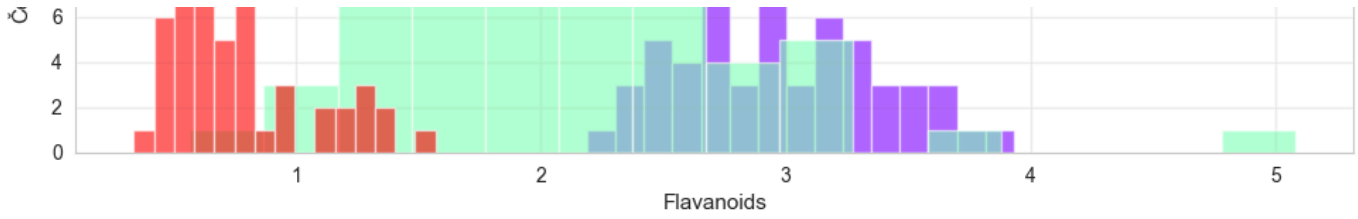


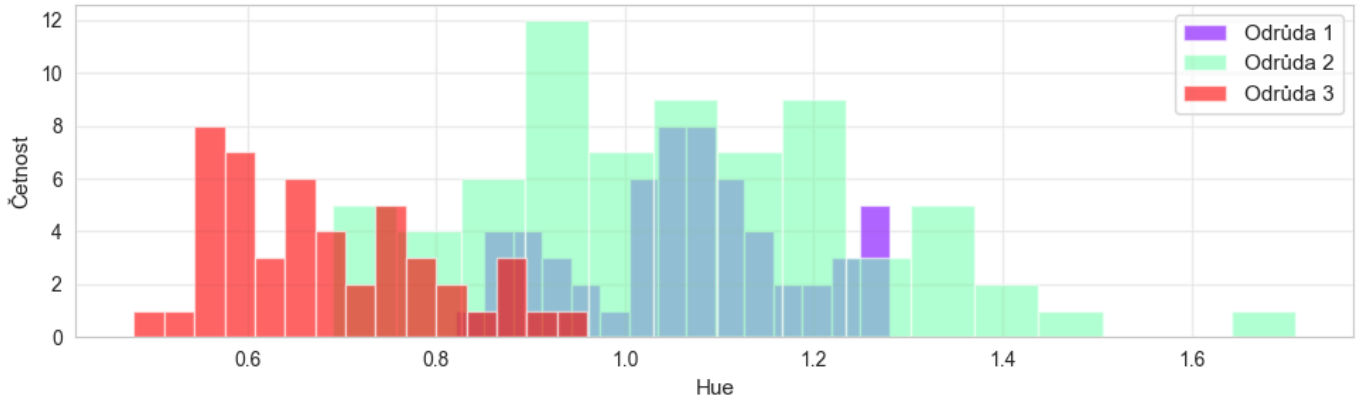


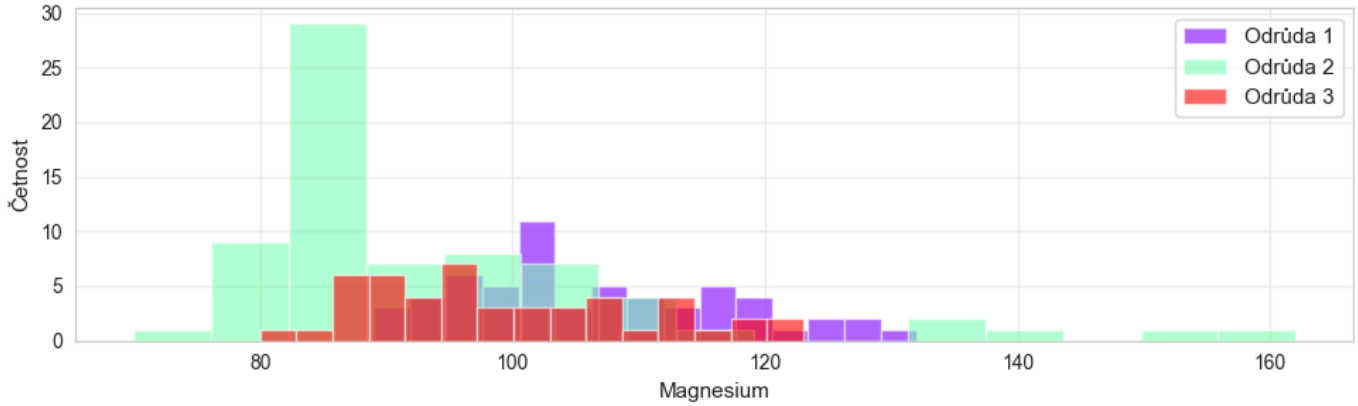


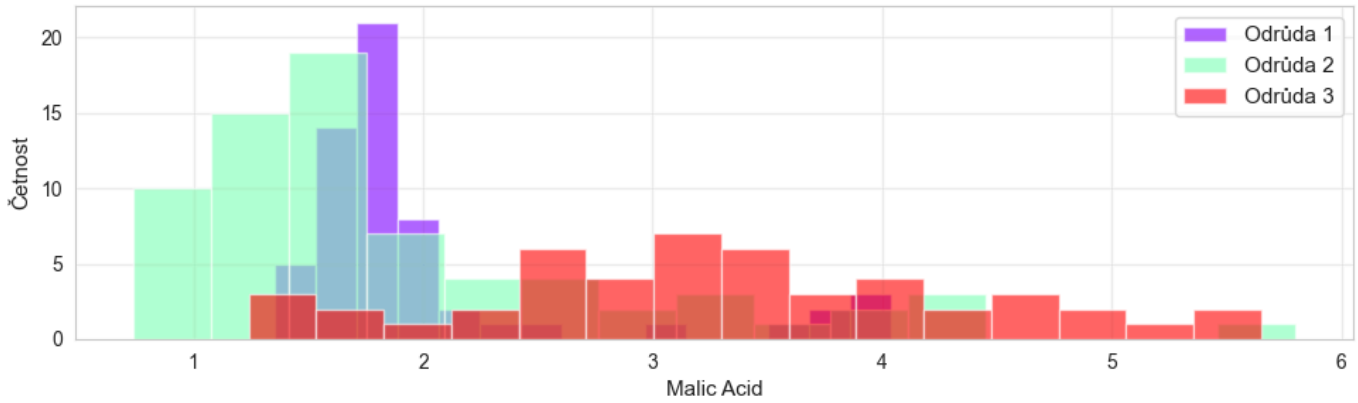


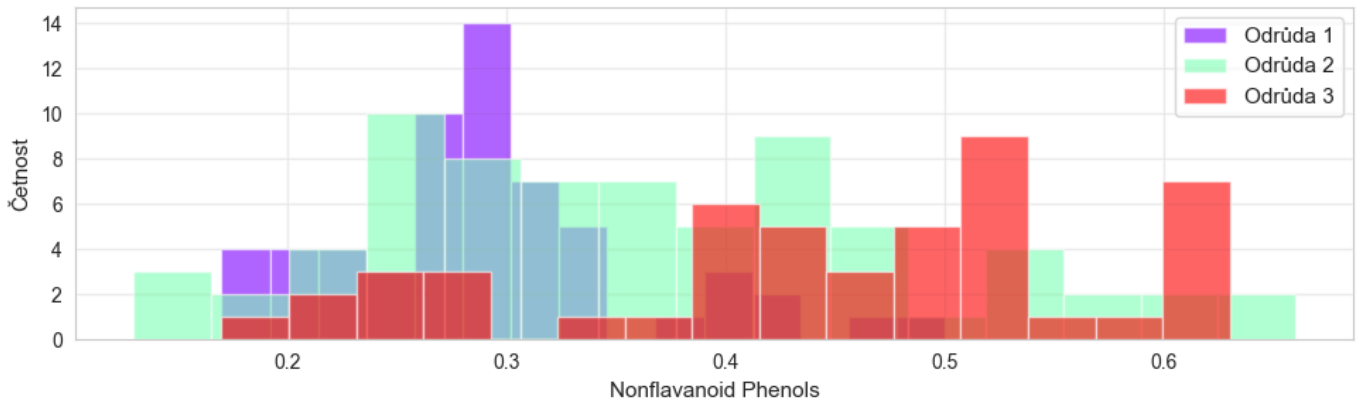


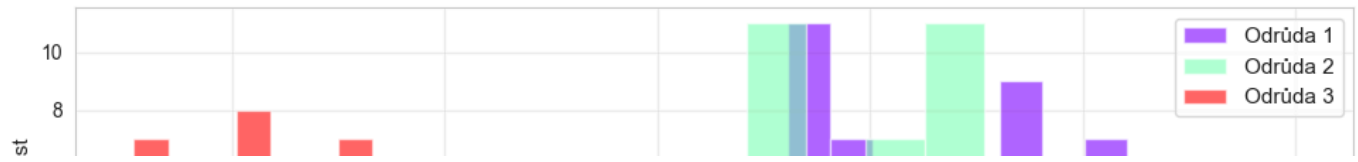

Flavanoids

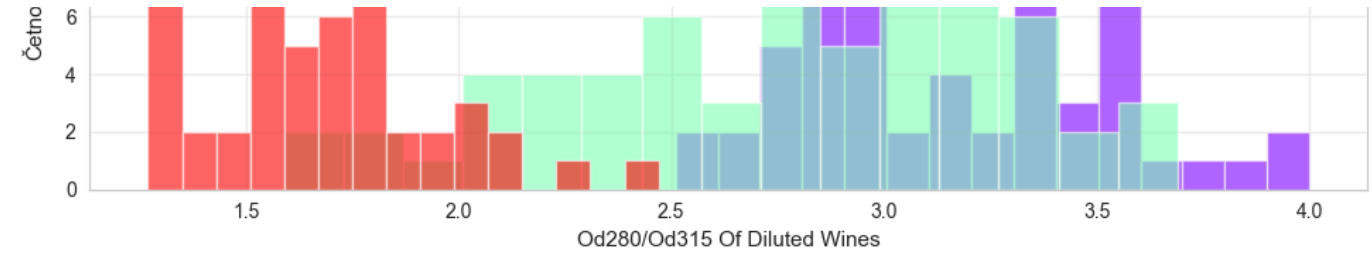

Četnost
Hue
Odrůda 1
Odrůda 2
Odrůda 3

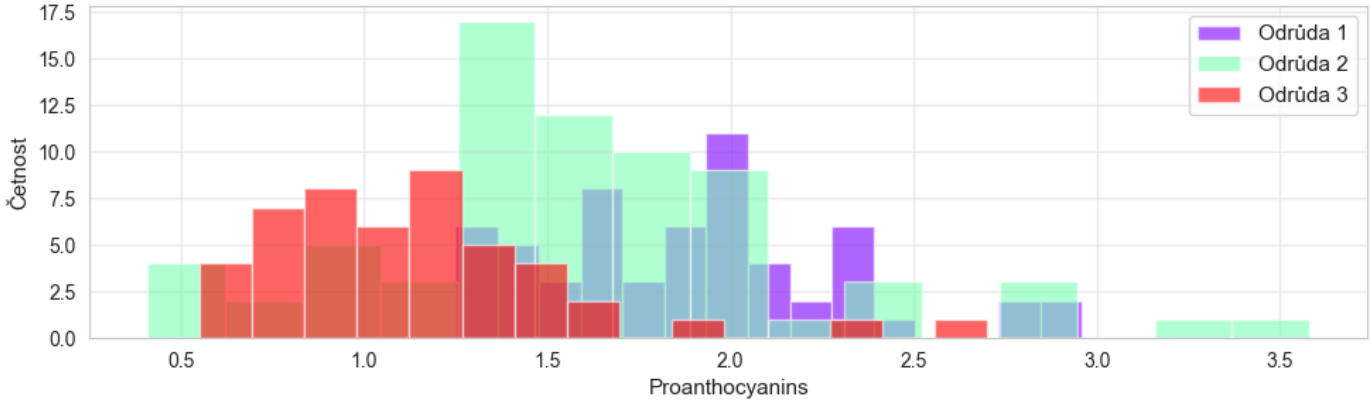

Četnost
Magnesium
Odrůda 1
Odrůda 2
Odrůda 3

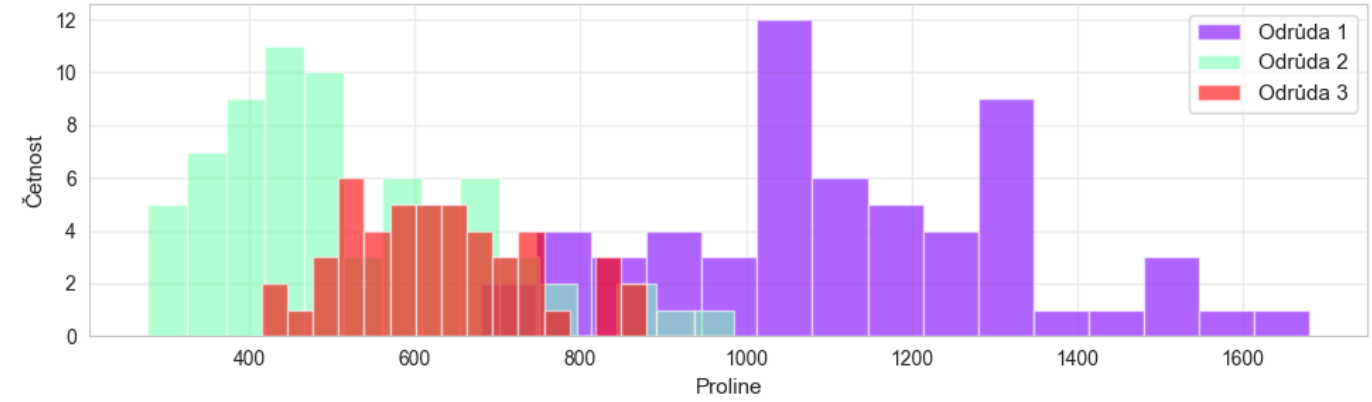

Četnost
Malic Acid
Odrůda 1
Odrůda 2
Odrůda 3

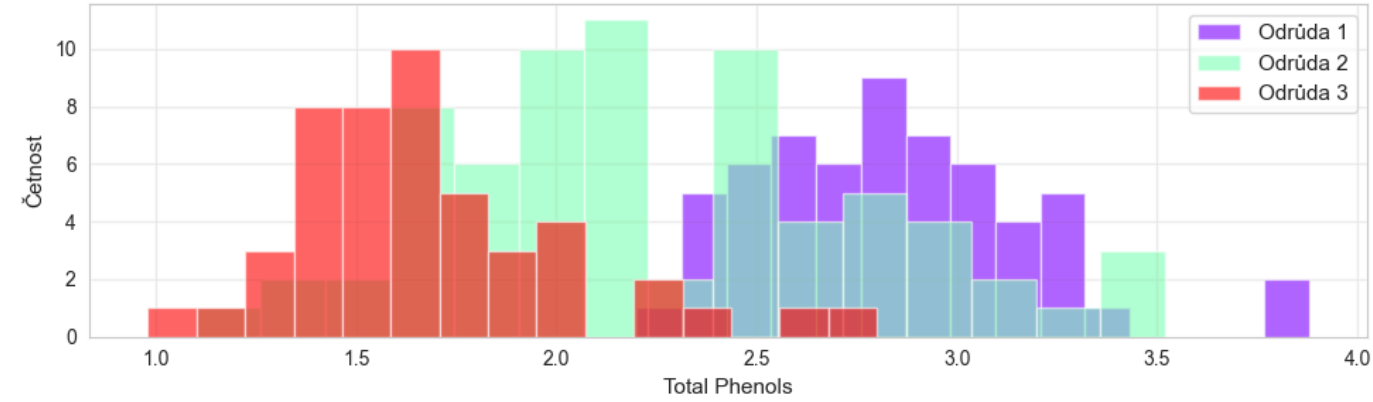

Četnost
Nonflavanoid Phenols
Odrůda 1
Odrůda 2
Odrůda 3


Odrůda 1
Odrůda 2
Odrůda 3









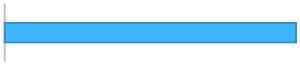
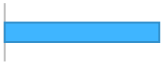
Předzpracování dat - Standardizace

Průměr po standardizaci (měl by být ~0): 0.000000
Směrodatná odchylka po standardizaci (měla by být ~1): 1.000000

Analýza hlavních komponent (PCA)

PCA redukuje dimenzionalitu dat transformací do nového prostoru hlavních komponent, které zachycují maximální variabilitu.

Vlastní čísla

Vlastní čísla			
PC	**Vlastní číslo**	**%**	**Kumulativní %**
1.00	**4.73**		**36**
2.00	**2.51**		**55**
3.00	**1.45**		**67**
4.00	0.92		74
5.00	0.86		80
6.00	0.65		85
7.00	0.55		89
8.00	0.35		92
9.00	0.29		94
10.00	0.25		96
11.00	0.23		98
12.00	0.17		99
13.00	0.10		100

Kaiserovo pravidlo (Kaiser Criterion)

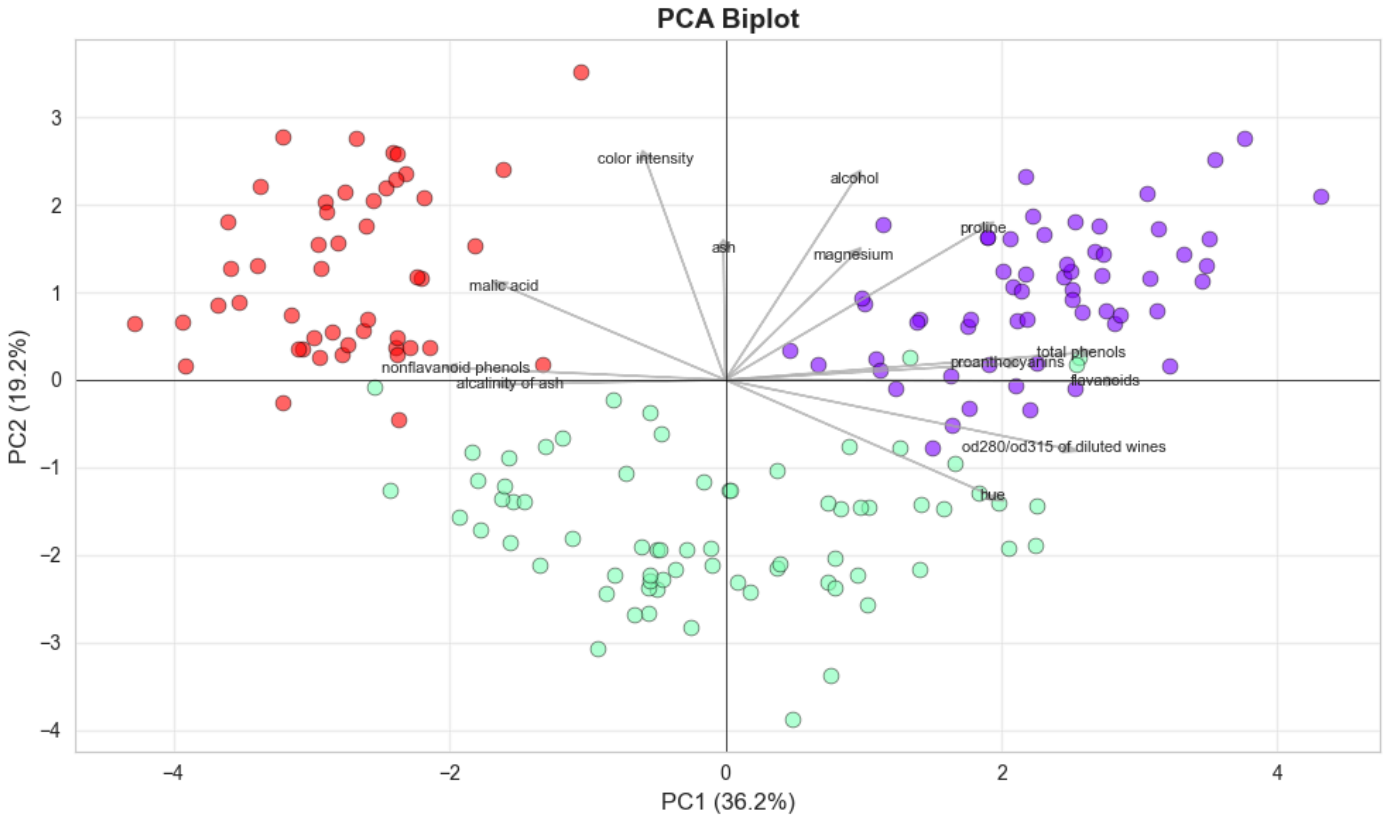
Používá se hlavně v případě, kdy je PCA prováděno na standardizovaných datech (korelační matice).

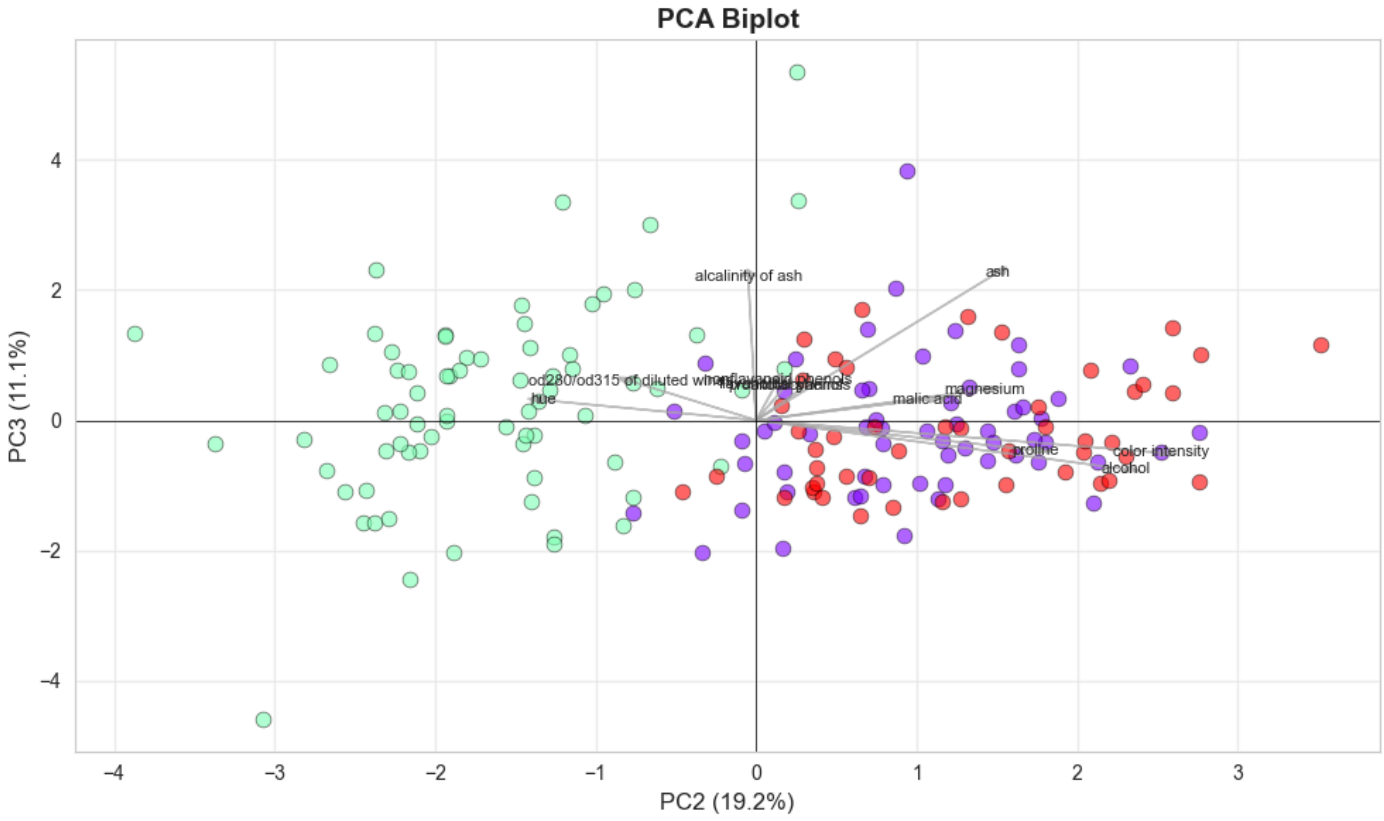
Pravidlo: Ponechat pouze ty PC, jejichž vlastní číslo $\lambda > 1$.

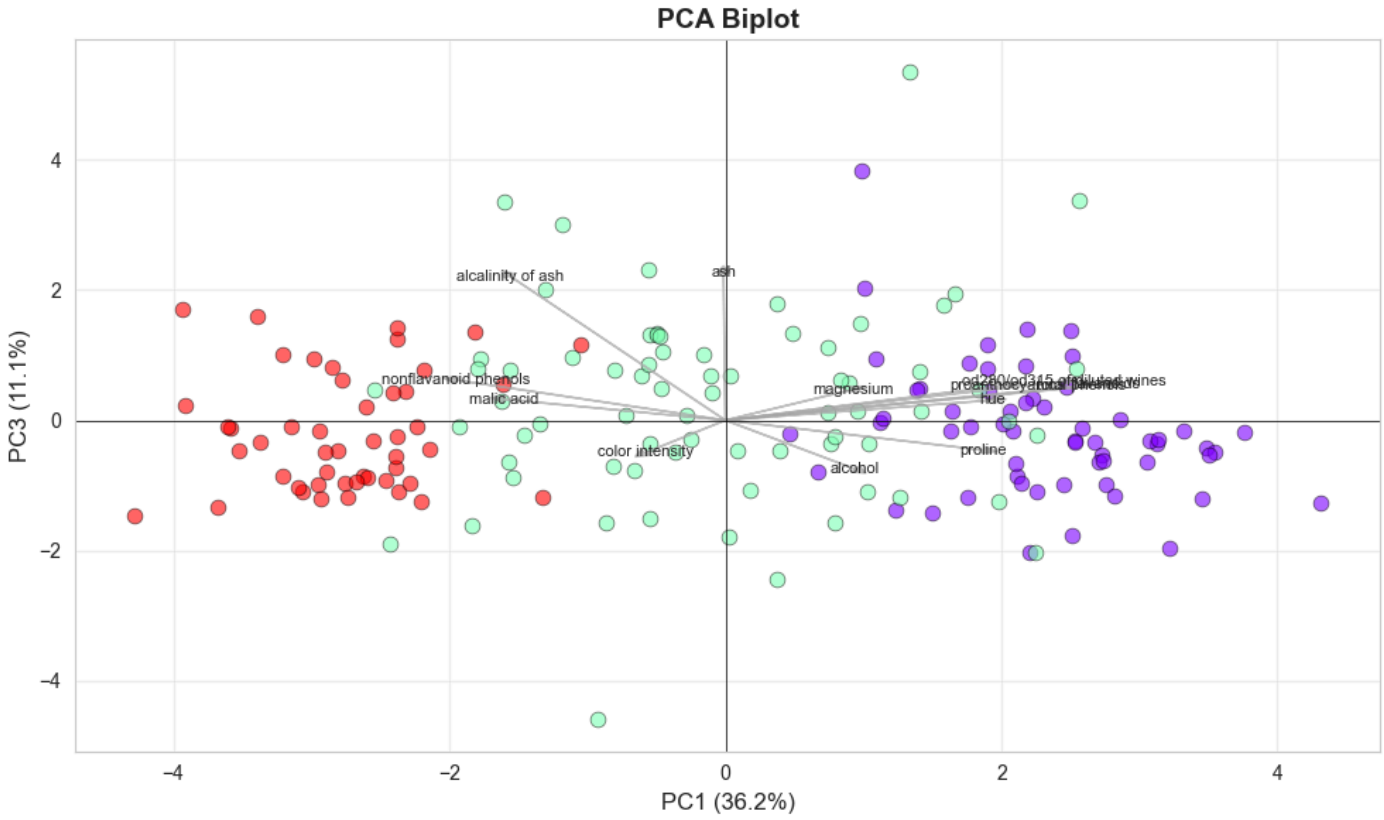
Proč: Vlastní číslo $\lambda = 1$ odpovídá rozptylu jedné původní proměnné. Pokud komponenta nevysvětlí ani tolik, co jedna původní proměnná, nemá smysl ji držet.

Vlastní vektory

Vlastní vektory						
Vlastnost	**PC1**	**PC1_%**	**PC1_%_CSUM**	**PC2**	**PC2_%**	**PC2_%_CSUM**
alcalinity_of_ash	−0.52	7	88	−0.02	0	100
alcohol	0.31	4	93	**0.77**	**17**	**36**
ash	−0.00	0	100	**0.50**	**11**	**60**
color_intensity	−0.19	3	100	**0.84**	**19**	**19**
flavanoids	**0.92**	**13**	**13**	−0.01	0	100
hue	0.65	9	65	−0.44	10	81
magnesium	0.31	4	97	0.47	11	71
malic_acid	−0.53	8	81	0.36	8	89
nonflavanoid_phenols	**−0.65**	**9**	**56**	0.05	1	100
od280/od315_of_diluted_wines	**0.82**	**12**	**37**	−0.26	6	95
proanthocyanins	**0.68**	**10**	**46**	0.06	1	98
proline	0.62	9	73	**0.58**	**13**	**49**
total_phenols	**0.86**	**12**	**25**	0.10	2	97


PCA Biplot
color intensity
alcohol
proline
ash
magnesium
malic acid
total phenols
proanthocyanins
nonflavanoid phenols
alcalinity of ash
flavanoids
od280/od315 of diluted wines
hue
PC2 (19.2%)
PC1 (36.2%)


PCA Biplot
alcalinity of ash
ash
od280/od315 of diluted wines
hue
malic acid
magnesium
proline
color intensity
alcohol
PC3 (11.1%)
PC2 (19.2%)



Interpretace PCA:

- První komponenta (PC1) vysvětluje **36,2% variability** a je silně korelována s fenolickými sloučeninami (flavonoidy, celkové fenoly)
- Druhá komponenta (PC2) vysvětluje **19,2% variability**
- Celkem první dvě komponenty zachycují **55,4% celkové variability**
- Odrůdy jsou částečně oddělitelné v PC prostoru
- PC3 je silně ovlivněna proměnnou Ash a Alcanility of Ash

Odlišení odrůd

- **Nejlepší pohled (PC1 vs PC2):**
  nejčistší rozdělení. PC1 a PC2 dohromady vysvětlují přes 55 % celkové variability dat, což je důvod, proč je tento pohled nejinformativnější.
- **Slabší separace (PC2 vs PC3):**
  velký překryv mezi Odrůdou 1 a 3. To znamená, že třetí komponenta (PC3) nepřidává mnoho informací pro jejich rozlišení.

Co charaketrizuje jednotlivé odrůdy

- **Odrůda 1:**
  Body ve směru šipek jako Proline, Phenols, Flavanoids a Alcohol. Tato vína mají tedy pravděpodobně nejvyšší obsah alkoholu a antioxidantů.
- **Odrůda 3:**
  Nachází se na opačné straně než většina šipek. To naznačuje, že tato vína mají nižší barevnost a méně flavonoidů. Jsou ve směru šipek Malic Acid a Nonflavanoid Phenols, což značí vyšší kyselost.
- **Odrůda 2:**
  V grafu PC1 vs PC2 směřuje k šipce Hue a od280/od315. To jsou vína s dobrou barevnou stabilitou.

Vztahy mezi proměnnými

- **Pozitivní korelace:**
  Šipky, které svírají malý úhel, spolu silně korelují. Pokud má víno hodně jednoho, má obvykle hodně i druhého, např. Flavanoids a Total Phenols
- **Negativní korelace:**
  Šipky mířící opačným směrem značí nepřímou úměru, např. Ash a Malic Acid v PC1 vs PC3
- **Nezávislost:**
  Šipky svírající pravý úhel spolu v těchto dimenzích nesouvisí.

Shrnutí:

Pro klasifikaci vín jsou nejdůležitější Flavanoids, Proline a Alcohol (PC1) a Color intensity/Hue (PC2). Třetí komponenta (PC3) je spíše doplňková a řeší detaily v minerálním složení (Ash).

Shluková analýza

K-means clustering

K-means pro více počtů clusterů

Metriky podle počtu clusterů		
Počet clusterů	**ARI**	**SIL**
2	0.37	0.27
3	**0.90**	**0.28**
4	0.82	0.25
5	0.63	0.18

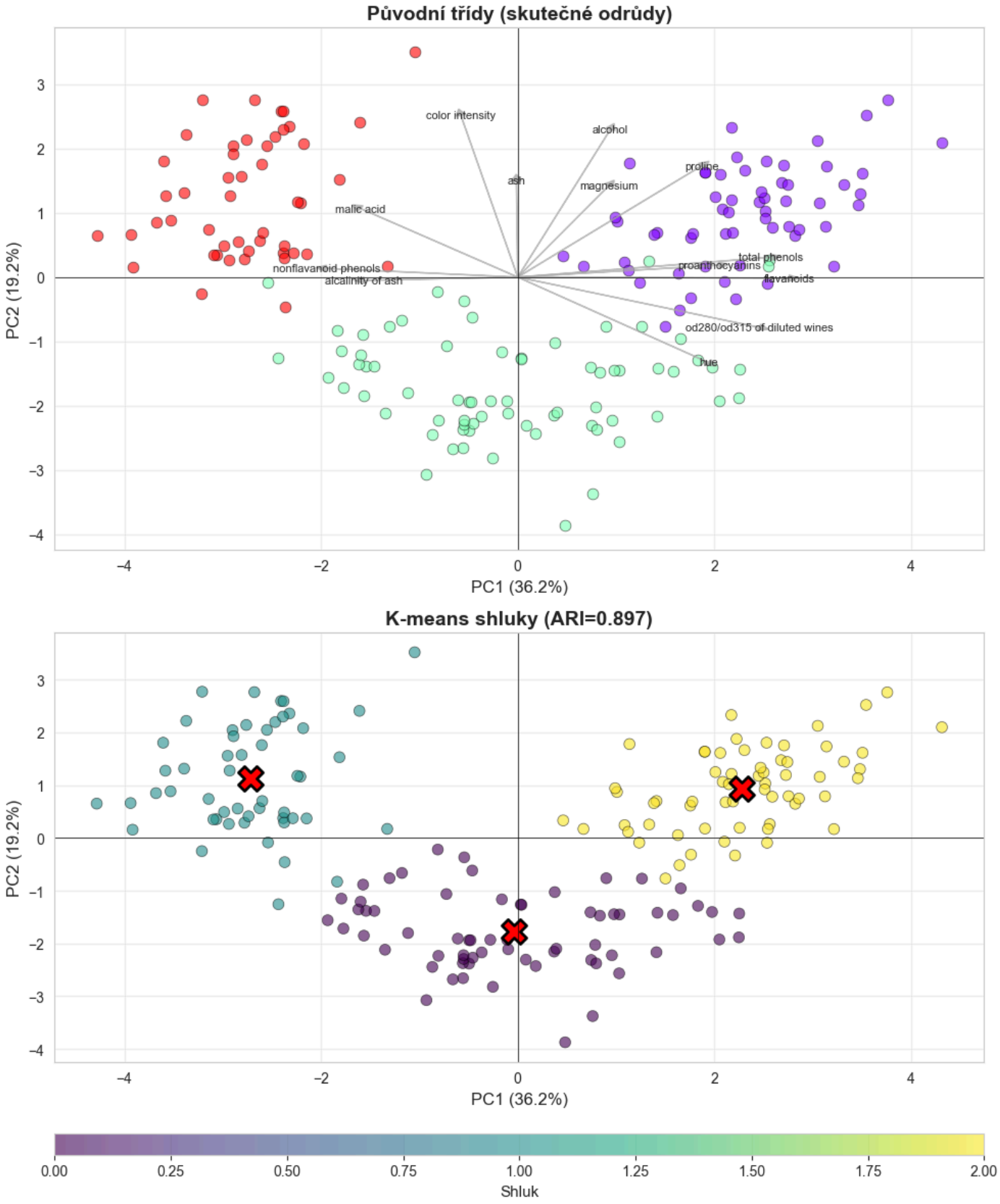
Metodika výběru počtu clasterů pro K-means

Pro určení optimálního počtu clusterů byla použita metoda grid-search, ze které při vyhodnocení ARI (Adjusted Rand Index) byl nejlepší výsledek s počtem 3, což se shoduje s počtem odrůd.

Interpretace K-means

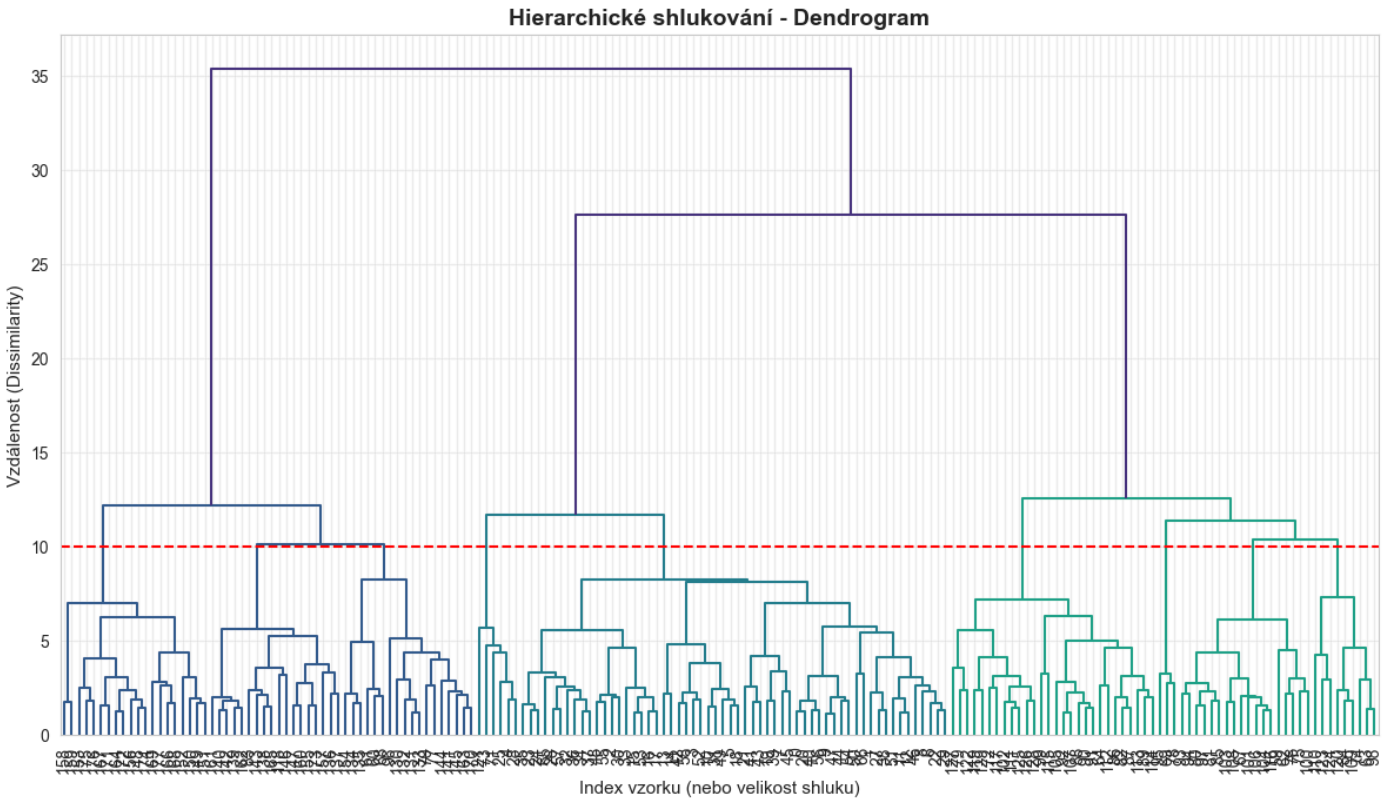
- **Vysoké ARI**:
  ukazuje na správné nalezení skupin, tedy že odrůdy byly identifikovány správně.
- **Nízká hodnota SIL (Silhouette score)**:
  separace není dobrá, tzn. že chemické vlastnosti odrůd se překrývají, což je přirozené a u dat biologického původu běžné, nebo je to důsledek vysoké dimenzionality, metoda K-means navíc předpokládá sférické clustery stejné velikosti, což nemusi struktuře dat odpovídat

Vizualizace K-means v PCA prostoru



Srovnání Shluků a Odrůd			
Predikovaný Shluk	**Odrůda 1**	**Odrůda 2**	**Odrůda 3**
0	0	65	0
1	0	3	48
2	59	3	0

Hierarchické shlukování

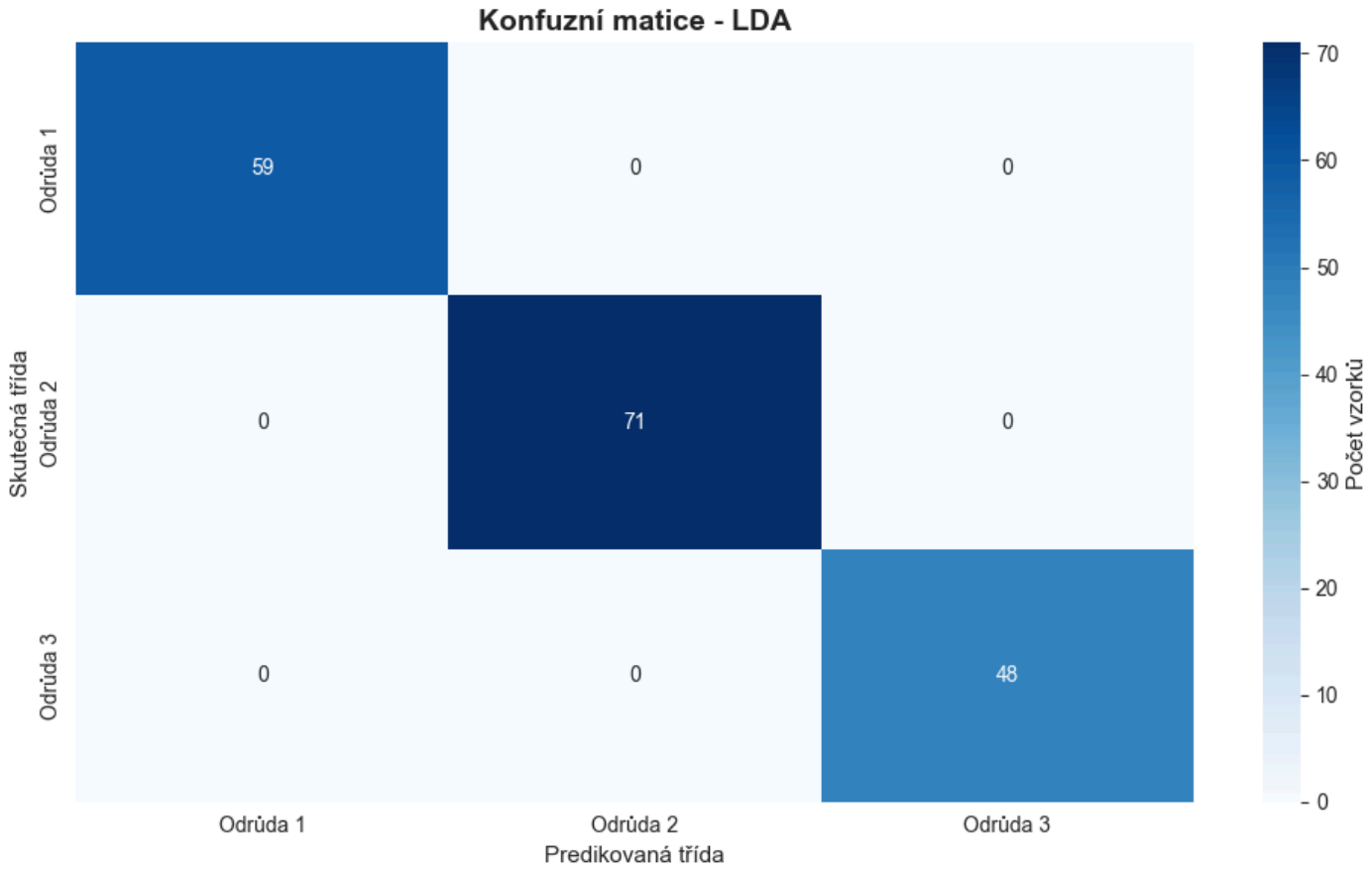


Srovnání Shluků a Odrůd			
Predikovaný Shluk	**Odrůda 1**	**Odrůda 2**	**Odrůda 3**
1	0	8	48
2	59	5	0
3	0	58	0

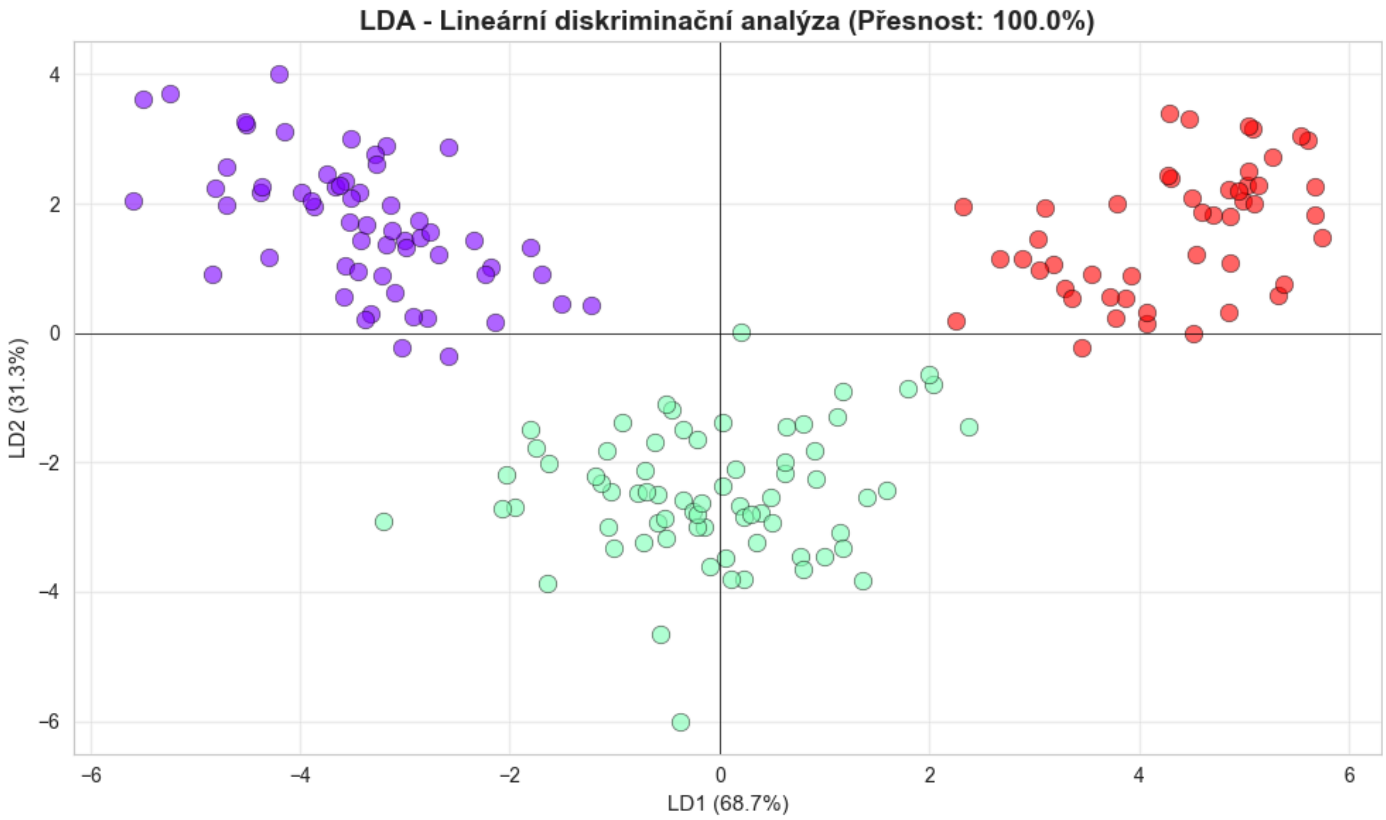
Lineární diskriminační analýza (LDA)

LDA hledá lineární kombinace proměnných, které maximálně oddělují předem definované třídy. Na rozdíl od PCA je to metoda řízená (supervised).

Přesnost klasifikace (LDA): 100.00%	
	Vysvětlený rozptyl diskriminačními funkcemi
LD1	0.69
LD2	0.31



Klasifikační Report				
Třída / Metrika	**Precision**	**Recall**	**F1-Score**	**Support**
Odrůda 1	1.000	1.000	1.000	59
Odrůda 2	1.000	1.000	1.000	71
Odrůda 3	1.000	1.000	1.000	48
accuracy	1.000	1.000	1.000	1
macro avg	1.000	1.000	1.000	178
weighted avg	1.000	1.000	1.000	178



Nejvýznamnější rysy pro LD1		
index	**LD1**	**LD2**
flavanoids	−1.65	−0.49
proline	−0.85	0.90
color_intensity	0.82	0.59
od280/od315_of_diluted_wines	−0.82	0.04
alcalinity_of_ash	0.52	−0.49

Nejvýznamnější rysy pro LD2		
index	**LD1**	**LD2**
proline	−0.85	0.90
alcohol	−0.33	0.71
ash	−0.10	0.64
color_intensity	0.82	0.59
flavanoids	−1.65	−0.49

Závěr a interpretace

Hlavní zjištění:

1. **PCA**
   - První dvě komponenty zachycují **55,4% variability**
   - Hlavní faktor variability: **fenolické sloučeniny** (flavonoidy, celkové fenoly)
   - Odrůdy jsou částečně oddělitelné v PC prostoru
2. **Shluková analýza**
   - K-means: Velmi vysoká shoda s původními třídami (**ARI = 0,897**), ale horší separabilita
   - Hierarchické shlukování potvrzuje existenci **3 přirozených shluků**
   - Data mají jasnou strukturu odpovídající třem odrůdám
3. **LDA (Lineární diskriminační analýza)**
   - Dosažena **100% přesnost klasifikace**
   - Odrůdy vín jsou **perfektně lineárně separabilní**

- LDA poskytuje nejlepší oddělení tříd (supervised metoda)

Praktický význam:

- Chemické vlastnosti umožňují **spolehlivou identifikaci odrůd vín**
- Metody vícerozměrné analýzy jsou účinným nástrojem pro:
  - Kontrolu kvality vín
  - Detekci falšování
  - Certifikaci původu
  - Charakterizaci odrůd

Doporučení:

- Pro **exploraci dat** → PCA
- Pro **hledání přirozených skupin** → Shluková analýza
- Pro **klasifikaci se známými třídami** → LDA
- V praxi: kombinace všech metod poskytuje nejkomplexnější pohled na data