PSM

Mnohorozměrné statistické metody

Úkol

Zadání

Úkoly

1. Je významný rozdíl mezi jednotlivými odrůdami v jejich chemickém složení? Ve které charakteristice se odrůdy nejvíce odlišují?

2. Pokuste se data zjednodušit, tj. najít několik málo nových proměnných, které když použiji namísto těch stávajících, tak ztratím co nejméně informace. Kolik takových proměnných potřebuji? A je možné je nějak interpretovat? Najděte rozumný počet nových proměnných, které nějakou interpretaci mít budou. Jejich počet se může lišit od doporučeného čísla. Kolik procent informace tyto proměnné obsahují?

3. Je možné na základě chemické analýzy vzorku vína poznat, ke které odrůdě víno patří? Určete vhodnou funkci / funkce, které Vám neznámý vzorek pomohou přiřadit ke správné skupině. Jak je taková rozhodovací funkce dobrá? A kolik takovýchto funkcí potřebujete?

4. V tomto případě pracujte bez znalosti odrůdy a jen na základě číselných proměnných rozdělte data do skupin. Porovnejte více variant dělení do skupin (různé metody, různé počty skupin, ...) a jedno dělení vyberte jako optimální. Kolik skupin máte? A jak byste tyto skupiny charakterizovali?

Data

Data pocházejí z databáze UCI a zaměřují se na italská vína.
Obsahují chemickou analýzu **178 vín** ze **3 odrůd** (proměnná Cultivar).
Níže naleznete popis jednotlivých proměnných:

Alcohol – Procentuální obsah alkoholu ve víně.
Malic.acid – Kyselina jablečná: druh kyseliny se silnou kyselostí a jablečným aroma. Červené víno je přirozeně doprovázeno kyselinou jablečnou.
Ash – Popel: podstatou popela je anorganická sůl, která ovlivňuje celkovou chuť vína a může mu dodat svěží pocit.
Alcalinity.of.ash – Alkalita popela: míra slabé zásaditosti rozpuštěné ve vodě.
Magnesium – Hořčík: prvek, který může podporovat energetický metabolismus a je slabě alkalický.
Total.phenols – Celkové fenoly: molekuly obsahující polyfenolické látky, které mají hořkou chuť a ovlivňují chuť, barvu a aroma vína. Patří k živinám ve víně.
Flavanoids – Flavanoidy: jsou antioxidantem prospěšným pro srdce a proti stárnutí, bohatým na aroma a hořkost.
Nonflavanoid.phenols – Neflavanoidní fenoly: speciální, slabě kyselé aromatické plyny s odolností proti oxidaci.
Proanthocyanin – Proanthokyany: bioflavonoidní sloučenina, která je přírodním antioxidantem s mírně hořkou vůní.
Color.intensity – Intenzita barvy: označuje stupeň barevného odstínu. Používá se k určení stylu vína, zda je „lehké“ nebo „husté“. Čím déle je víno během procesu výroby v kontaktu s hroznovou šťávou, tím hustší je chuť a intenzivnější barva.
Hue – Odstín: označuje sytost a „teplo“ barvy. Lze jej použít k určení odrůdy a stáří vína. Červená vína s vyšším stářím budou mít žlutý odstín a zvýšenou průhlednost. Intenzita barvy a odstín jsou důležitými ukazateli pro hodnocení kvality vzhledu vína.
OD280.OD315.of.diluted.wines – OD280/OD315 zředěných vín: metoda pro stanovení koncentrace bílkovin, která umožňuje určit obsah bílkovin v různých vínech.
Proline – Prolin: hlavní aminokyselina v červeném víně a důležitá součást nutričních a chuťových hodnot vína.

Řešení

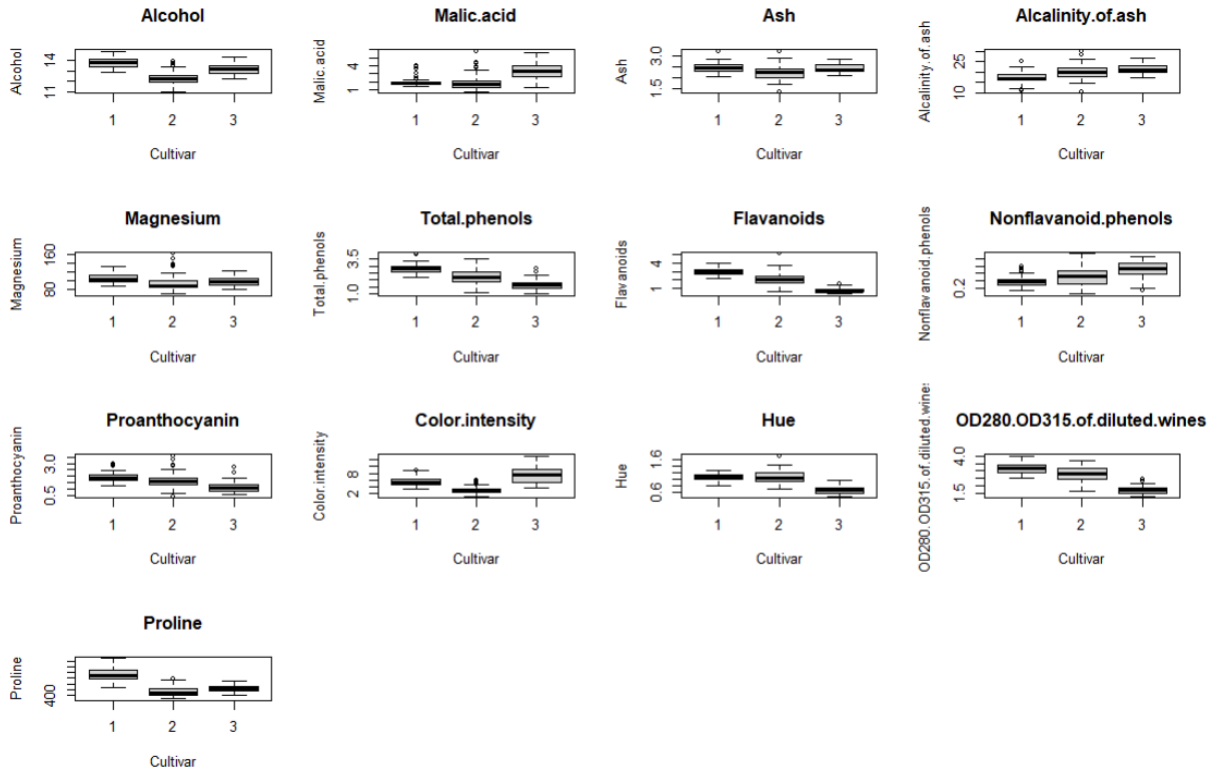
Úkol 1

Zadání

Je významný rozdíl mezi jednotlivými odrůdami v jejich chemickém složení?
Ve které charakteristice se odrůdy nejvíce odlišují?

Řešení

Nejdříve jsem se podíval na box-ploty skupin v rámci všech proměnných:



Už podle tohoto bych řekl, že by mohl být významný rozdíl ve většině proměnných.
S MANOVA testuji nulovou hypotézu, že jsou všechny 3 skupiny stejné napříč všemi sloupci.
Vyšla mi p-hodnota 2.2e-16, díky čemuž zamítám nulovou hypotézu a potvrzují že jsou vskutku rozdílné a tato informace je statisticky významná.

V summary MANOVY jsem zjistil, že se skupiny liší ve všech proměnných velmi významným způsobem, ovšem v některých víc než v jiných - porovnávám pomocí F-hodnoty:

1. Flavanoids: F value = 233.93
2. Proline: F value = 207.92
3. OD280.OD315.of.diluted.wines: F value = 189.97
4. Alcohol: F value = 135.08
5. Color.intensity: F value = 120.66
6. Hue: F value = 101.32
7. Total.phenols: F value = 93.733
8. Malic.acid: F value = 36.943
9. Alcalinity.of.ash: F value = 35.772
10. Proanthocyanin: F value = 30.271
11. Nonflavanoid.phenols: F value = 27.575
12. Magnesium: F value = 12.43
13. Ash: F value = 13.313

První tři hodnoty mají velmi vysoké hodnoty, pak hodnoty 4 - 7 mají taky podobně vysoké a zbytek spadne velmi nízko oproti zbytku. Troufnu si říct, že nejvíc se odrůdy odlišují ve **Flavanoids**, **Proline** a **OD280.OD315.of.diluted.wines**.

Kód

```
########################################  ÚKOL 1
########################################
# Je významný rozdíl mezi jednotlivými odrůdami v jejich chemickém složení?
# Ve které charakteristice se odrůdy nejvíce odlišují?

pairs(data)
# total.phenols X Flavanoids = silný lineární vztah


numeric_data <- data[, names(data) != "Cultivar"]
n <- ncol(numeric_data)
rows <- ceiling(sqrt(n))
cols <- ceiling(n / rows)
par(mfrow = c(rows, cols))
for (var in names(numeric_data)) {
  boxplot(numeric_data[[var]] ~ data$Cultivar,
          main = var,
          xlab = "Cultivar", ylab = var)
}
# Které proměnné mají vlic na rozdělení skupin okometricky
# Alcohol
# Malic.acid ale jen pro 3, ostatni jsou stejny
# Total.phenois a Flavanoids = hodně podobné, zároveň mají silně lineární vztah. nejspíš se
mohu jedné z nich zbavit
# nonflavanoid phenols podobně jako flavanoids ale inverted
# OD280... 3 je vytazne mensi
# Proline
# Hue 3 je výrazně menší

manova_model <- manova(as.matrix(wine[, -1]) ~ as.factor(wine$Cultivar))
summary(manova_model, test = "Wilks")
# Jo, je rozdíl mezi skupinami -> 2.2e-16 p-hodnota
# Zamítáme nulovou hypotézu, že neexistuje žádný rozdíl v průměrných vektorech
# chemických složení mezi odrůdami. Závěr je, že existuje statisticky významný
# rozdíl v celkovém chemickém složení mezi jednotlivými odrůdami vína.

aov_summary <- summary.aov(manova_model)
# Odrůdy se nejvíce odlišují v charakteristikách Flavanoids, Proline, Alcohol, Color.intensity
a OD280.OD315.of.diluted.wines
```

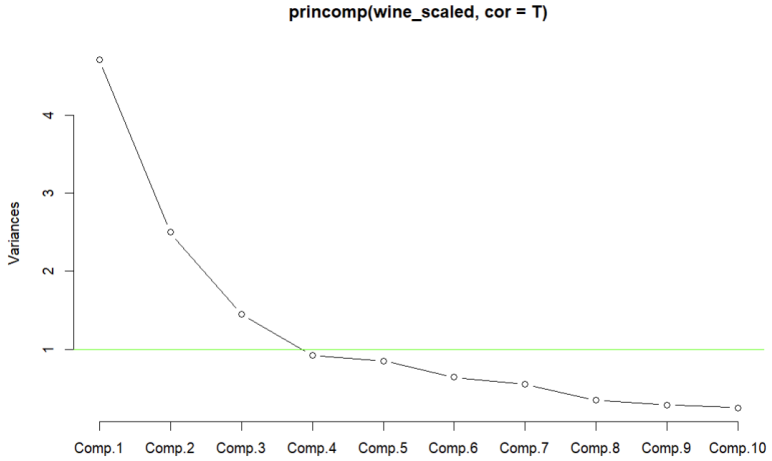
Úkol 2

Zadání

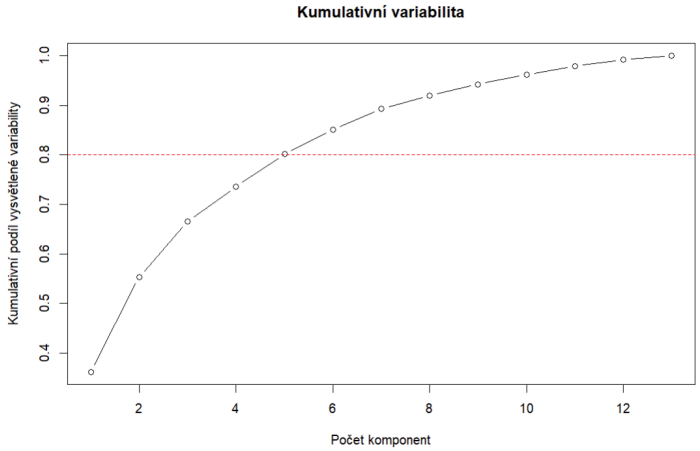
Pokuste se data zjednodušit, tj. najít několik málo nových proměnných, které když použiji namísto těch stávajících, tak ztratím co nejméně informace.
Kolik takových proměnných potřebuji?
A je možné je nějak interpretovat?
Najděte rozumný počet nových proměnných, které nějakou interpretaci mít budou.
Jejich počet se může lišit od doporučeného čísla.
Kolik procent informace tyto proměnné obsahují?

Řešení

Původně jsem využil metodu PCA.
Pomocí scree-plotu jsem došel k tomu, že bych měl použít 3-4 komponenty i podle Kaiserového kritérie i podle toho, že se tam láme loket (nebo jak se tomu správně říká :D)



Poté jsem si vykreslil graf ukazující vysvětlenou variabilitu dat komponentami. Řekl jsem si, že 80% vysvětlené variability bude stačit - z toho jsem došel k tomu, že prvních 5 komponent bude stačit. 20% ztracené informace není až tak špatné.



Získal jsem tedy 5 komponent, která mi nově vysvětlují data, ale protože nedojdu k nijak výživné interpretaci, přešel jsem k faktorové analýze.

Opět s pomocí scree-plot jsem se rozhodl pro 3 faktory (4. faktor nedával moc smysl).

```
Loadings:
                        Factor1 Factor2 Factor3
Alcohol                           0.779
Malic.acid              -0.470
Ash                                       0.629
Alcalinity.of.ash                         0.856
Magnesium
Total.phenols            0.824
Flavanoids               0.928
Nonflavanoid.phenols    -0.533
Proanthocyanin           0.622
Color.intensity         -0.413   0.748
Hue                      0.654
OD280.OD315.of.diluted.wines  0.864
Proline                           0.688
```

Vyšly mi tyto faktory:
poznámka: nerozumím tak důkladně vínu a už vůbec ne chemickému složení, takže jsem si s interpretací nechal pomoc od všem tak známé umělé inteligence, výsledky od ní jsem zvýraznil fialově.

- Faktor 1
  - Zachycuje strukturu fenolických látek (flavonoidy, antioxidanty) a jejich vliv na barvu a chuť.
  - Faktor by se tedy mohl nazvat **Barva a chuť**
- Faktor 2
  - Vysoké hodnoty alkoholu, prolinu (aminokyselina) a barevné intenzity → něco o plnosti těla vína / intenzitě chuti.
  - Faktor by se tedy mohl nazvat **Intenzita a plnost**
- Faktor 3
  - Složení popela a alkalita říkají něco o minerálním profilu vína.
  - Faktor by se tedy mohl nazvat **Minerální profil**

Z vlastní vůle a zvědavosti (a rozhodně ne zadání) jsem se rozhodl prozkoumat kolik procent variability tyto faktory vysvětlují.

	Faktor	Variabilita	Kumulativně
Factor1	1	30.84	30.84
Factor2	2	17.52	48.36
Factor3	3	10.02	58.38

Dohromady vysvětlují 58% variability. To není zrovna moc. Ale po menším pátrání jsem došel k tomu, že na vysvětlené variabilitě u FA moc nezáleží. Hlavní je aby samotné faktory dávaly smysl v kontextu a něco doopravdy vysvětlovaly.

Kód

```
######################################  ÚKOL 2
######################################
# Pokuste se data zjednodušit,
# tj. najít několik málo nových proměnných, které když použiji namísto těch stávajících, tak ztratím co nejméně informace.
# Kolik takových proměnných potřebuji?
# A je možné je nějak interpretovat?
# Najděte rozumný počet nových proměnných, které nějakou interpretaci mít budou.
# Jejich počet se může lišit od doporučeného čísla.
# Kolik procent informace tyto proměnné obsahují?

# Standardizace (PCA je citlivá na měřítko)
wine_scaled <- scale(wine_numeric)

pca_model <- prcomp(wine_scaled, scale = TRUE)
summary(pca_model)

# Nejspíš 3 nebo 4 komponenty
screeplot(princomp(wine_scaled, cor = T), type="l")
abline(h=1, col="green")

explained_var <- pca_model$sdev^2 / sum(pca_model$sdev^2)
cum_var <- cumsum(explained_var)

# Tabulka nebo graf
plot(cum_var, type = "b", xlab = "Počet komponent", ylab = "Kumulativní podíl vysvětlené variability", main = "Kumulativní variabilita")
abline(h = 0.8, col = "red", lty = 2)  # 80% hranice
# řekl bych 5 komponent, protože vysvětlují 80% variability dat (ztratím 20% informace)
# PCA skóre pro jednotlivá pozorování
PC_scores <- pca_model$x
# Ze všech sloupců jsem si vytvořil nové faktory které mi vysvětlí data v méně sloupcích
# Ale protože je to PCA tak nemůžu dobře odhadnout co ty faktory zastupují (kombinace jakých sloupců to je)
# Proto využiju Faktorovou analýzu



# Faktorová analýza
screeplot(princomp(wine_scaled, cor = T), type="l")
abline(h=1, col="green")
# řekl bych 3 faktory
fa_model <- factanal(wine_scaled, factors = 3, rotation = "varimax", scores = "Bartlett")
# Faktorové loadingy
# Loading je číslo, které říká jak silně daná proměnná „patří“ k danému faktoru.
# Důvod proč je faktorová analýza gut - můžu říct co ten faktor zastupuje
```

```
# Hodnoty nad 0.4 -> přispívá to
# +/- = směr vztahu
print(fa_model$loadings, cutoff = 0.4)  # hodnoty pod 0.4 potlač
# Nerozumím vínu tak jsem si nechal vysvětlil GPT co ty hodnoty znamenají a na co by mohly ukazovat.
# Faktor 1 = zachycuje strukturu fenolických látek (flavonoidy, antioxidanty) a jejich vliv na barvu a chuť.
# Faktor 2 = Vysoké hodnoty alkoholu, prolinu (aminokyselina) a barevné intenzity → něco o plnosti těla vína / intenzitě chuti.
# Faktor 3 = Složení popela a alkalita říkají něco o minerálním profilu vína.

# Pokud ty faktory dávají logický smysl tak by se nemusela řešit variabilita - ve FA jde o to jestli to dává smysl v kontextu
# Stejně si vypočítám ještě kolik variability to vysvětluje... mohlo by stačit nad 60-70%

# Kolik procent variability mi ty faktory vysvětlují?
ss_loadings <- colSums(fa_model$loadings^2)
# Celková vysvětlená variabilita (v %)
explained_variance <- ss_loadings / ncol(wine_scaled)
cumulative_variance <- cumsum(explained_variance)

data.frame(
  Faktor = 1:length(ss_loadings),
  Variabilita = round(explained_variance * 100, 2),
  Kumulativně = round(cumulative_variance * 100, 2)
)
# Dohromady vysvětlí 58%
# To je málo ale neřeším protože faktory dávají smysl!
```

Úkol 3

Zadání

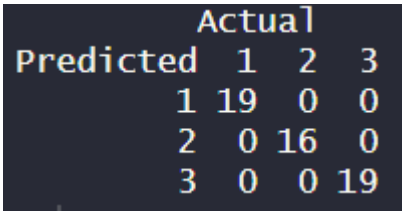
Je možné na základě chemické analýzy vzorku vína poznat, ke které odrůdě víno patří?
Určete vhodnou funkci / funkce, které Vám neznámý vzorek pomohou přiřadit ke správné skupině.
Jak je taková rozhodovací funkce dobrá?
A kolik takových funkcí potřebujete?

Řešení

“Je možné na základě chemické analýzy vzorku vína poznat, ke které odrůdě víno patří? “

- Ano.

Rozhodl jsem se využít diskriminační analýzu (protože nic jiného neznám :-)), přesněji lineární diskriminační analýzu.
Potřebuji dva lineární diskriminanty, protože rozděluji do tří skupin.
Model LDA jsem nechal natrénovat na 70% dat a otestoval na zbylých 30%.



	Actual		
Predicted	1	2	3
1	19	0	0
2	0	16	0
3	0	0	19

Model předpovídá se 100% přesností.

Kód

```
########################################  ÚKOL 3
########################################
# Je možné na základě chemické analýzy vzorku vína poznat, ke které odrůdě víno patří?
# Určete vhodnou funkci / funkce, které Vám neznámý vzorek pomohou přiřadit ke správné skupině.
# Jak je taková rozhodovací funkce dobrá?
# A kolik takových funkcí potřebujete?


if (!require(MASS)) install.packages("MASS")
library(MASS)

wine_data <- scale(wine[, -1])  # standardizace

# Rozdělení dat na train a test
set.seed(543)
train_idx <- sample(1:nrow(wine), 0.7 * nrow(wine))
train_data <- wine[train_idx, ]
test_data <- wine[-train_idx, ]

# Trénování modelu
lda_model <- lda(Cultivar ~ ., data = train_data)

# Výsledky
lda_model
plot(lda_model)  # Vizualizace oddělení skupin podle LD1 a LD2

# Predikce a hodnocení
pred <- predict(lda_model, newdata = test_data)
# Matice záměn (confusion matrix): ukazuje, kolik vzorků bylo zařazeno správně / špatně.
table(Predicted = pred$class, Actual = test_data$Cultivar)
# Přesnost klasifikace: % správných klasifikací.
mean(pred$class == test_data$Cultivar)

# 100% funkční!
```

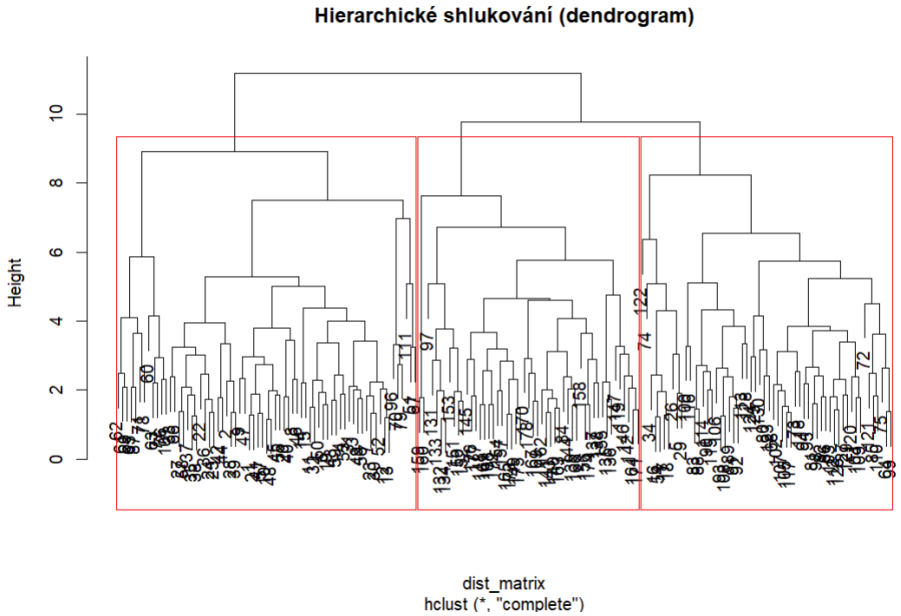
Úkol 4

Zadání

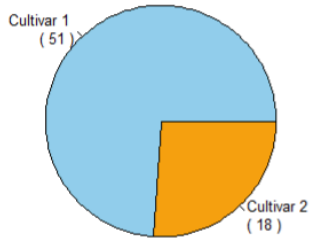
V tomto případě pracujte bez znalosti odrůdy a jen na základě číselných proměnných rozdělte data do skupin.
Porovnejte více variant dělení do skupin (různé metody, různé počty skupin, ...) a jedno dělení vyberte jako optimální.
Kolik skupin máte?
A jak byste tyto skupiny charakterizovali?

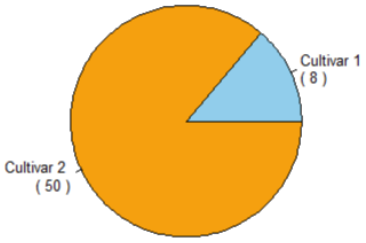
Řešení

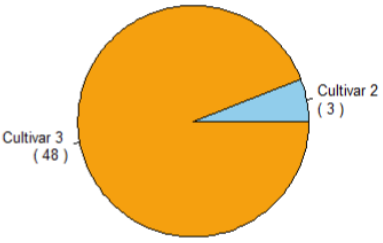
Jako první jsem zkusil hierarchické shlukování s dendrogramem.
Vyzkoušel jsem metody “complete”, “average” a “ward.D2.



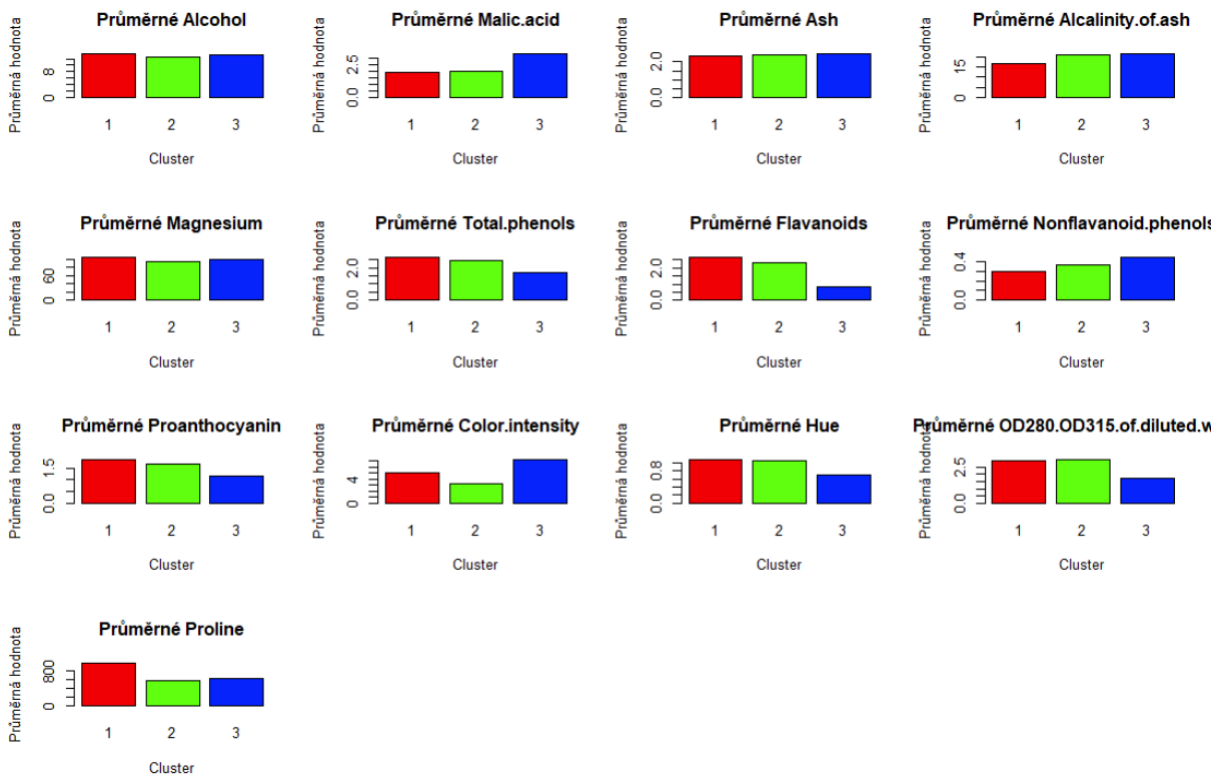
Z nich se mi nejvíc líbila metoda complete, která vychází nejlépe na 3 skupiny (stejně jako ward.D2). Kvůli počtu skupin jsem vyzkoušel jestli na to náhodou nemá vliv odrůda (i přesto, že mám pracovat bez znalosti odrůdy…) a zdá se, že to bude tím.
(Koláčový graf zleva do prava Cluster1, 2 a 3)







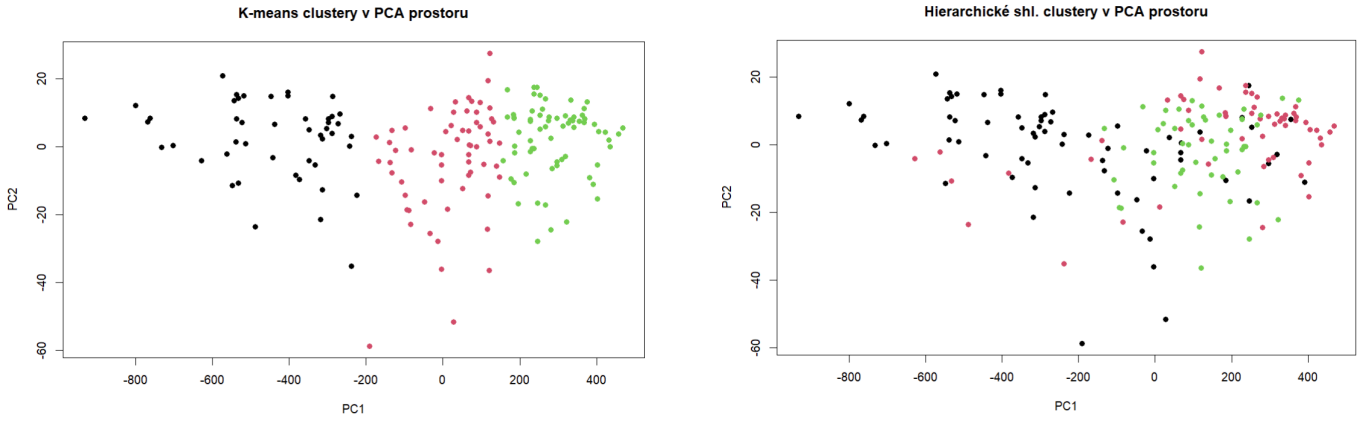
Na grafech jde vidět že se skupiny liší primárně odrůdami. Ovšem kvůli zadání se podíváme čím jiným by to mohlo být.



Na grafu jde vidět průměrné hodnoty různých chemikálií pro každý cluster.
Podle mého názoru jsou největší rozdíly v:

- Color intensity
- Flavanoids
- Proline
- HUE
- OD280…

Vyzkoušíme ještě k-means clustering.



K-means se ukázal jako o hodně spolehlivější metoda oproti hierarchickému shlukování.

```
> kmeans_model$centers
   Alcohol Malic.acid      Ash Alcalinity.of.ash Magnesium Total.phenols Flavanoids Nonflavanoid.phenols Proanthocyanin
1 13.80447   1.883404 2.426170          17.02340 105.51064      2.867234   3.014255             0.2853191       1.910426
2 12.92984   2.504032 2.408065          19.89032 103.59677      2.111129   1.584032             0.3883871       1.503387
3 12.51667   2.494203 2.288551          20.82319  92.34783      2.070725   1.758406             0.3901449       1.451884
  Color.intensity       Hue OD280.OD315.of.diluted.wines   Proline
1        5.702553 1.0782979                     3.114043 1195.1489
2        5.650323 0.8839677                     2.365484  728.3387
3        4.086957 0.9411594                     2.490725  458.2319
```

Největší vliv na clustery k-means mají:

- Proline
- Alcanity of ash
- Magnesium
- Flavanoids
- Color intensity

Kód

```
######################################  ÚKOL 4
######################################
# V tomto případě pracujte bez znalosti odrůdy a jen na základě číselných proměnných rozdělte data do skupin.
# Porovnejte více variant dělení do skupin (různé metody, různé počty skupin, ...) a jedno dělení vyberte jako optimální.
# Kolik skupin máte?
# A jak byste tyto skupiny charakterizovali?

### Hned zkusím dendrogram protože ho mám rád

distance_matrix <- dist(wine)
# Použijeme hierarchické shlukování (agregativní metoda: "complete")
hc <- hclust(dist_matrix, method = "complete") # Využít tenhle, rozděluje to tak nějak podle
hc  <- hclust(dist_matrix, method = "average") # NOPE
hc     <- hclust(dist_matrix, method = "ward.D2")  # Asi nejhezčí
#- hclust() provede samotné shlukování, v tomto případě s metodou "complete",
# což znamená, že dva shluky se spojí, když nejvíce vzdálené objekty mezi nimi mají co nejmenší vzdálenost.
# Vykreslíme dendrogram
plot(hc, main = "Hierarchické shlukování (dendrogram)")
k_val <- 3
seg <- cutree(hc, k = k_val)
# rozdeli data do 3 skupin
rect.hclust(hc, k=k_val, border="red")
# Jaké sloupce tam jsou?

### Vidím že 3 clustery jsou docela ideální, že by to bylo rozdělením skupin?

# 2. Přidáme informaci o shluku zpět do dat
data_temp <- wine
data_temp$Cluster <- seg
# 3. Tabulka poměrů Cultivarů v každém shluku
table_list <- split(data_temp$Cultivar, data_temp$Cluster)
counts <- lapply(table_list, table)
# 4. Vykreslíme 3 koláčové grafy vedle sebe
par(mfrow = c(1, k_val))  # rozložení 1 řádek, 3 sloupce

for (i in 1:k_val) {
  pie(
    counts[[i]],
    main = paste("Cluster", i),
    col = c("skyblue", "orange", "seagreen"),
    labels = paste("Cultivar", names(counts[[i]]), "\n(", counts[[i]], ")")
  )
```

```
}
par(mfrow = c(1, 1))

### Skupiny za to mohou částečně, v první clusteru je pouze 1. odrůda a ve zbylých dvou jsou víceméně napůl 2. a 3.
### V prnvím úkolu jsem přišel na to že se skupiny nejvíc liší ve Flavanoids, Proline a OD280...
### Třeba je to tím ?
vars <- setdiff(names(wine), c("Cultivar"))
# Nastavíme grafické okno: dynamicky podle počtu proměnných
par(mfrow = c(4, 4))

# Automaticky vykreslíme barplot pro každou proměnnou
for (v in vars) {
  means <- tapply(wine[[v]], as.factor(seg), mean)
  barplot(
    means,
    main = paste("Průměrné", v),
    xlab = "Cluster",
    ylab = "Průměrná hodnota",
    col = rainbow(k_val)
  )
}

# Reset grafického rozložení
par(mfrow = c(1, 1))


# K-means clustering
require(graphics)
# pocet skupin beru na zaklade predesleho hierarchickeho shlukovani
kmeans_model <- kmeans(wine[, !(names(wine) == "Cultivar")], centers = k_val, nstart = 25)
table(seg, kmeans_model$cluster)
# Jak vypadají shluky pro každou dvojici proměnných
pairs(wine, col = kmeans_model$cluster, pch = 19)
# Vizualizace v PC
# „PC“ znamená Principal Components = hlavní komponenty z metody PCA (Principal Component Analysis).
# Rozdělí data na dvě hlavní komponenty a nad nima kontroluji rozdíly
pc <- prcomp(wine)
plot(pc$x[, 1], pc$x[, 2], col = kmeans_model$cluster, pch = 19,
     main = "K-means clustery v PCA prostoru", xlab = "PC1", ylab = "PC2")
# Porovnání s hierarchickým shlukováním
plot(pc$x[, 1], pc$x[, 2], col = seg, pch = 19,
     main = "Hierarchické shl. clustery v PCA prostoru", xlab = "PC1", ylab = "PC2")
# Co má vliv na K-means clustery?
kmeans_model$centers
# Proline — rozdíl přes 700, nejvíce se liší, silně ovlivňuje rozdělení.
```

Alcalinity.of.ash — rozdíl skoro 4, hodně významné.
Magnesium — rozdíl přes 13, velký vliv.
Flavanoids — rozdíl přes 1.4, důležitý faktor.
Color.intensity — rozdíl přes 1.6, důležitý.
Total.phenols, OD280.OD315.of.diluted.wines, Alcohol — střední rozdíly, taky přispívají.
Ostatní proměnné mají menší rozdíly, tedy menší vliv.