Úkol č. 3

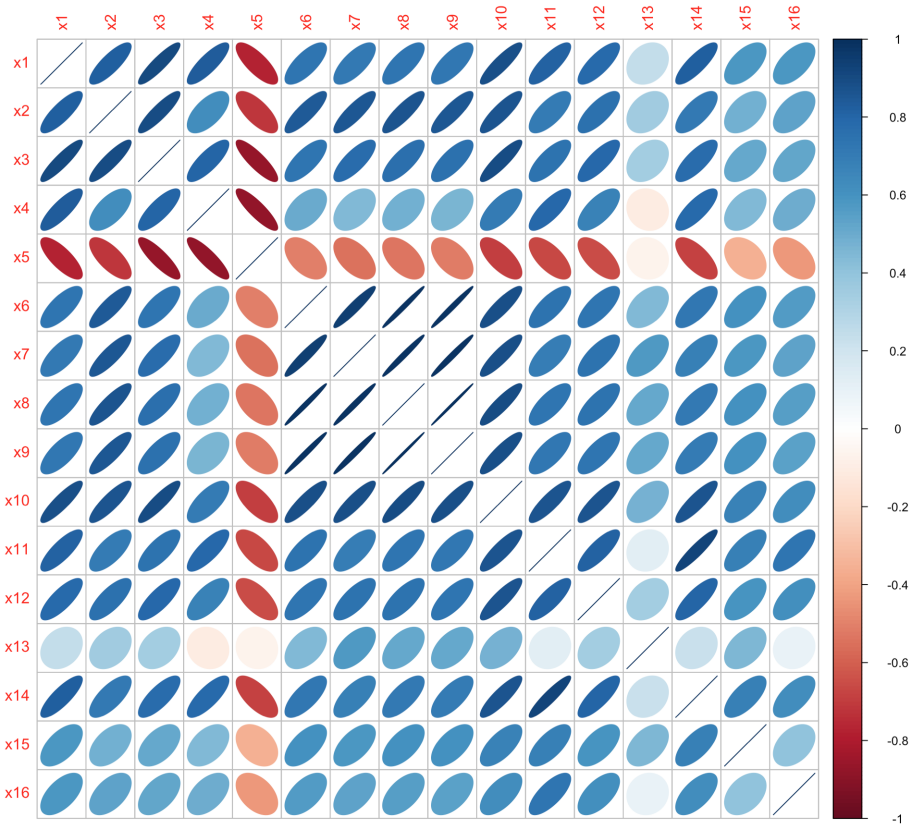
Jaroslav Bezděk

ZS 2017/2018

Obsah

1 Výběr proměnných pro shlukovou analýzu

Proměnné jsem vybíral tak, že jsem se podíval na schéma korelací jednotlivých proměnných, viz obrázek č. 1. Na tomto schématu jsem hledal takové proměnné, které silně korelují s jednou či několika dalšími proměnnými. Takové proměnné jsem postupně ze souboru vyřazoval, protože lze říct, že při zanechání obou proměnných by v souboru byl jistý druh informace reprezentován dvakrát.



Obrázek 1: Korelace pro jednotlivé proměnné, výstup z R.

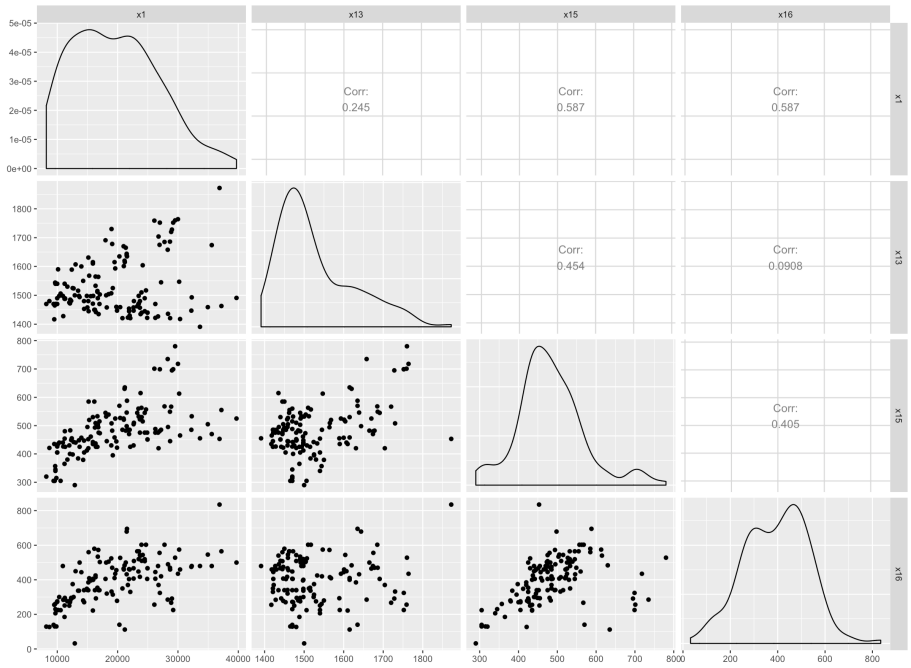
Na základě této úvahy jsem ze souboru vyřadil proměnné *x6*, *x7* a *x9*, které silně korelují s proměnnou *x8*. Tuto proměnnou, která uvádí kombinovanou spotřebu, jsem v souboru ponechal.

Další proměnnou, se kterou silně korelují jiné proměnné, je proměnná *x1*. Na základě silné korelace s touto proměnnou jsem se souboru vyřadil proměnné *x2*, *x3*, *x4*, *x5*, *x8*, *x10*, *x11*, *x12* a *x14*.

V datovém souboru jsem tedy ponechal čtyři proměnné: *x1*, *x13*, *x15* a *x16*.

2 Průzkumová analýza dat

Při průzkumové analýze jsem zjistil, že tři ze čtyř proměnných mají spíše kladně zešikmené rozdělení a většina proměnných spolu slabě až středně silně koreluje. Z obrázku z č. 2 je také patrné, že některé proměnné mají malý počet odlehlých pozorování.



Obrázek 2: Přehled vybraných proměnných, výstup z R.

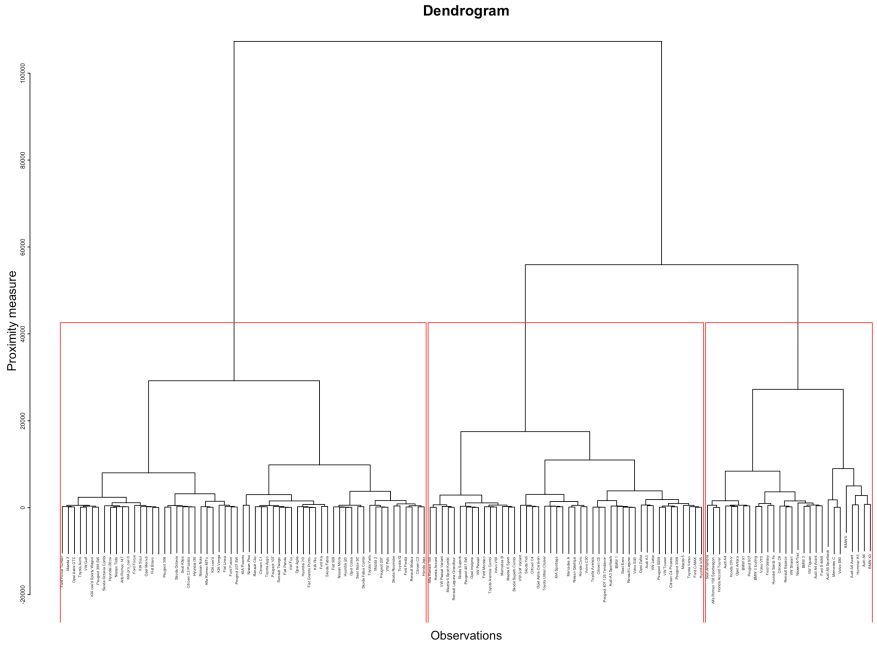
Co se týče odlehlých pozorování, myslím si, že soubor obsahuje i jednu chybu. Jedná se o záznam č. 114, tedy model *Toyota iQ*. I přestože se jedná o velice malý automobil, data uvádí, že objem jeho zavazadlového prostoru je 32 l. Tato hodnota je o jeden řád menší, než hodnota u automobilu s druhým nejmenším objemem zavazadlového prostoru (112 l). Abych mohl se záznamem dále pracovat, našel jsem na internetu[1] správnou hodnotu a záznam opravil.

Dále jsem zjistil, že v souboru se nachází jedna duplicita. Konkrétně se jedná o automobil modelu *Honda Accord*. Ten se v souboru nachází dvakrát. Duplicitní záznam jsem z datového souboru vyřadil.

[1] https://www.vybermiauto.cz/katalog/toyota/aygo-5dv/technicka-data

3 Shluková analýza

Snažil jsem se vytvořit přibližně stejně velké shluky, proto jsem pro shlukování vybral Wardovu metodu. Jak ukazuje obrázek č. 3, tento záměr se přibližně podařilo splnit.



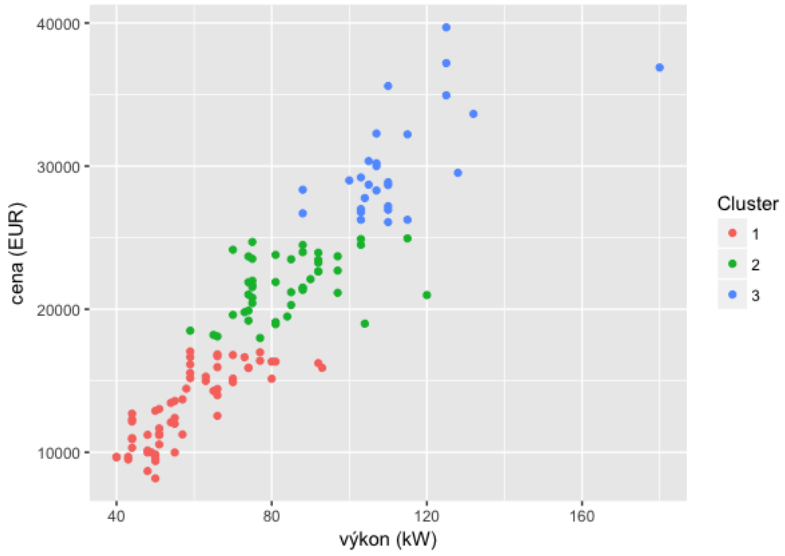
Obrázek 3: Dendrogram se zvýrazněnými třemi výslednými shluky, výstup z R.

Nejrozumnější mi přijde vytvořit tři shluky automobilů. Tyto shluky jsou charakterizovány rozdílnými cenami, velikostí i výkonem automobilu. Rozdílné hladiny těchto proměnných pro různé shluky ukazují obrázky č. 4 i a č. 5.

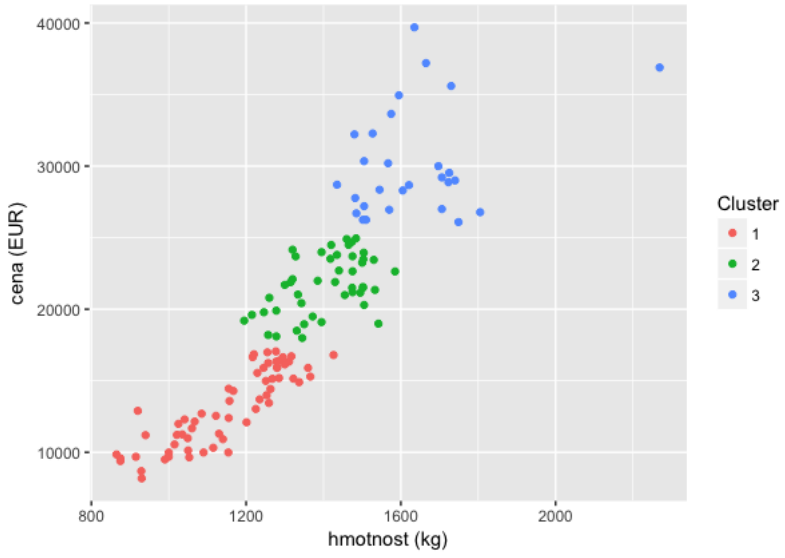
Z nich je patrné, že do prvního shluku byly zařazeny automobily s nejnižší cenou, výkonem i hmotností. Jinými slovy se tedy jedná o nejlevnější automobily, které jsou výkonově slabé a ze sledovaného souboru jsou také nejmenší. Tento shluk by tedy bylo možné označit jako automobily nižší třídy.

Do druhého shluku byly zařazeny automobily vyšší ceny, které zároveň mají větší výkon i hmotnost. Nemají však tak vysoké hodnoty těchto proměnných jako automobily třetího shluku. Druhý shluk tedy dle mého názoru obsahuje automobily střední třídy.

Ve třetím shluku se nacházejí automobily s nejvyšším výkonem, cenou i hmotností. Tento shluk tedy obsahuje velká a silná auto, která nejsou cenově tak dostupná jako auta z prvního a druhého shluku. Dle mého názoru se tedy jedná o automobily vyšší třídy.



Obrázek 4: Závislost ceny automobilu a výkonu automobilu, výstup z R.



Obrázek 5: Závislost ceny automobilu a hmotnosti utomobilu, výstup z R.