Úkol 3 - Shluková analýza

4ST512 Vícerozměrná statistika

Jiří Filip

Úvod

Cílém této práce je provést shlukovou analýzu na datech, jež obsahují údaje (např. cena, objem válců, výkon, maximální rychlost) o 136 modelech automobilů. Data obsahují pouze vozy vybraných značek do cenového limitu 40 000 EUR a každý model je zastoupen vždy nejlevnějším typem v nabídce.

Naším cílem bude vybrat minimálně 3 a maximálně 12 proměnných k identifikaci tržních segmentů. Výběr těchto proměnných bude důkladně zdůvodněn a popsán. K segmentaci použijeme shlukovou analýzu a rovněž popíšeme, proč jsme zvolil zrovna tu danou metodu.

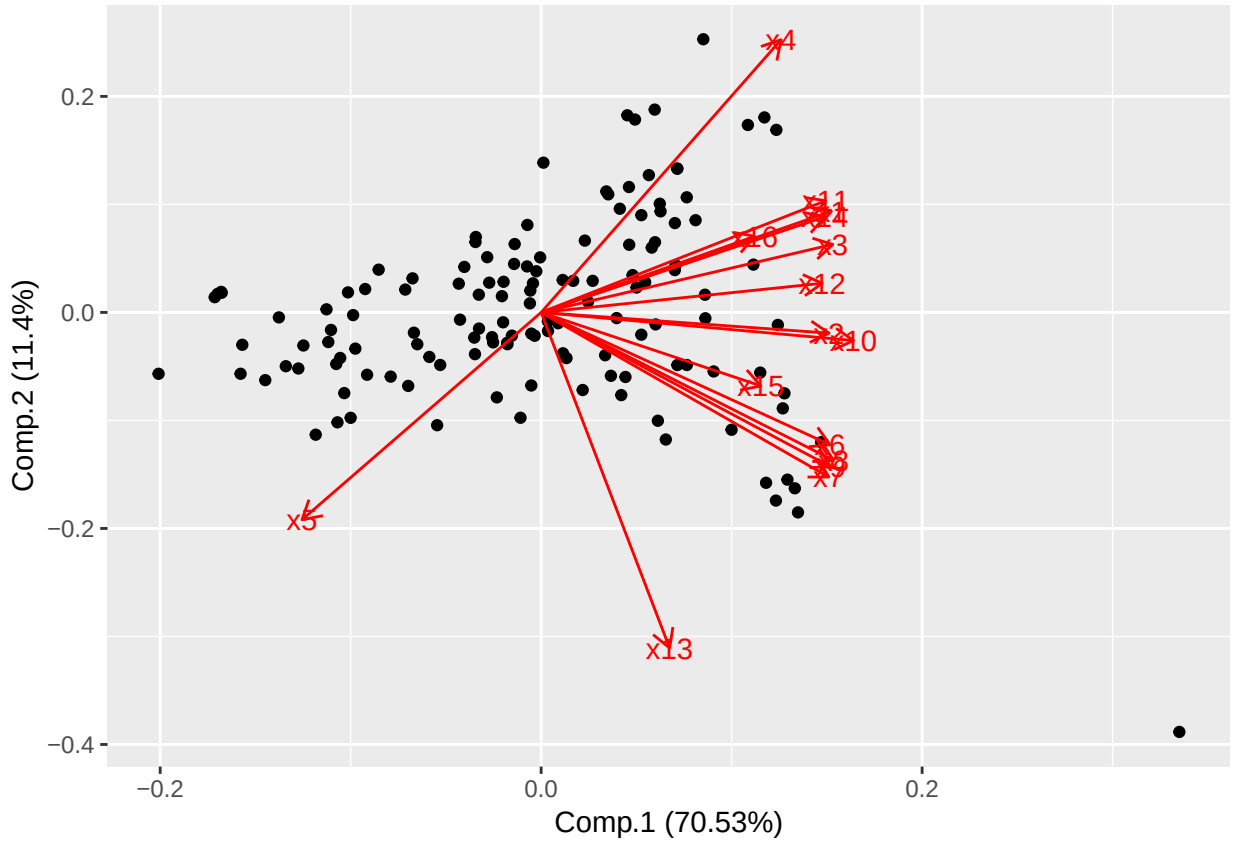
Výsledné shluky interpretujeme a popíšeme kroky, kterými se nám podařilo je najít.

Průzkum dat

K průzkumu dat využijeme metodu hlavních komponent. Této metody využijeme zejména proto, že později budeme zobrazovat výsledné shluky na projekci proměnných do podprostoru prvních dvou hlavních komponent.

Na grafu můžeme vidět, že vpravo v dolním rohu je jedno odlehlé pozorování. Jde o pozorování č. 45 (Hummer H3). Z dat ho odstraníme.

Z grafu dále vidíme, že pro vysvětlení variability dat by byla vhodná jedna z proměnných x5 (zrychlení) nebo x4 (vektory zátěží napovídají, že jsou negativně korelované), dále proměnná x13 (výška) a rozumně vybrané další proměnné, jejichž vektory zátěží nejsou příliš blízko sebe (což by znamenalo korelaci).

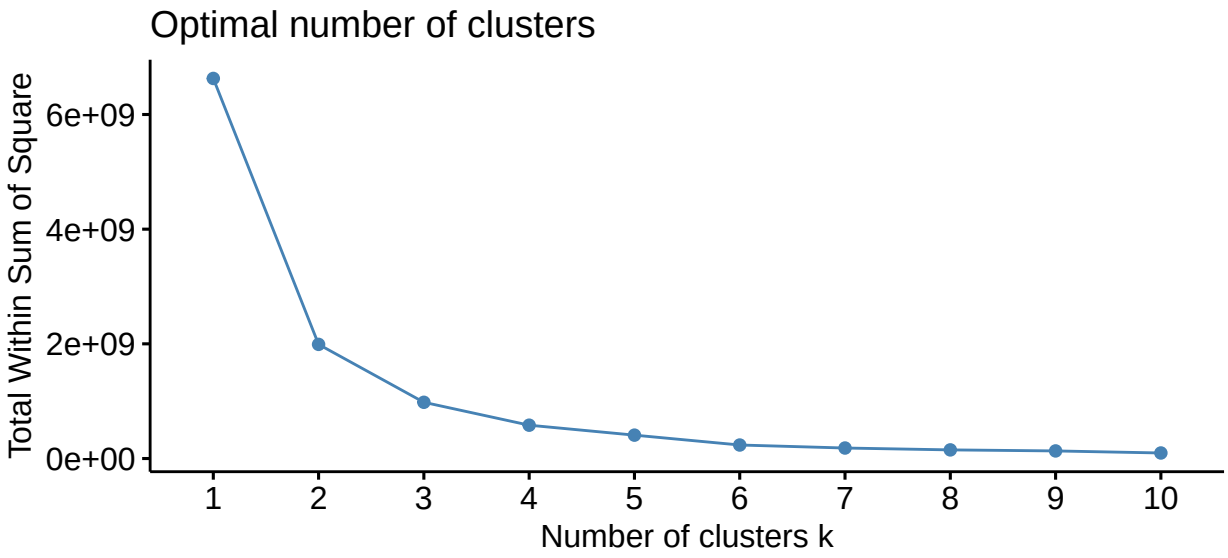


Určení počtu shluků

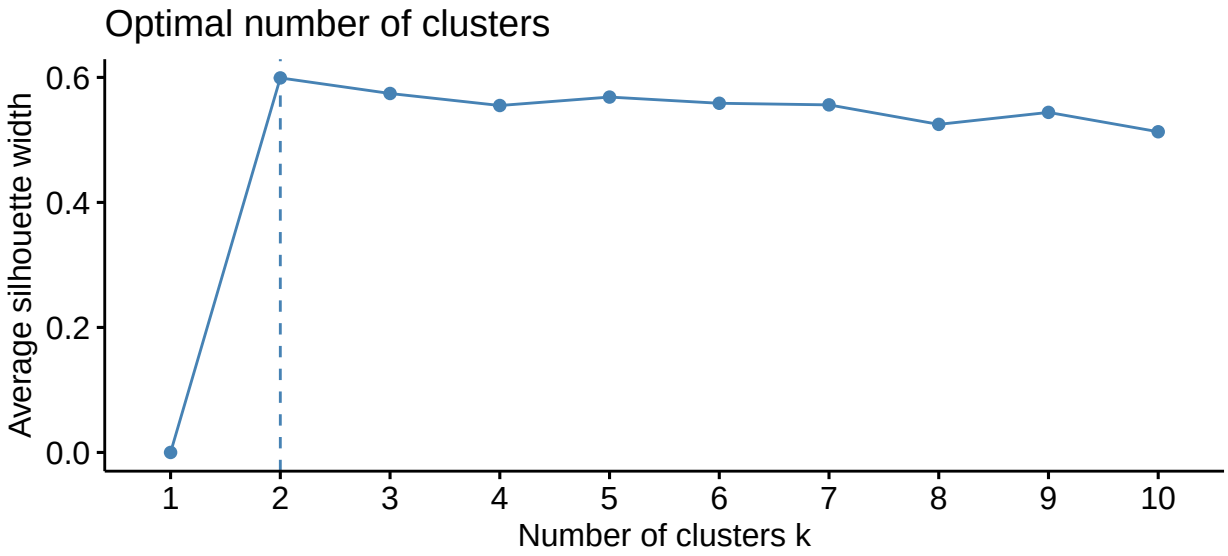
Nakonec jsme se rozhodli pro využití nehierarchického shlukování a algoritmu K-means. Toto rozhodnutí plyne z toho, že máme spíše zájem o provedení tržní segmentace na nějaký počet shluků než provedení rozpadu (nebo spojení) segmentů. Určitě by bylo zajímavé pokusit se interpretovat, jak je nějaký tržní segment tvořen jinými, menšími segmenty, ale pro účely jednoduchosti použijeme algoritmus K-means s určeným počtem shluků.

Algoritmus K-means nalezne lokální optimum, proto algoritmus spustíme několikrát, abychom si byli jisti, že jsme nalezli uspokojivé řešení.

Nyní zbývá se rozhodnout, kolik shluků necháme vytvořit. K tomu nám dopomůžou dva grafy. Jeden z nich zobrazuje součet čtverců mezishlukových vzdáleností (které se snažíme minimalizovat). Vídíme pak, že graf se “láme” přibližně na 3 - 5 shlucích.

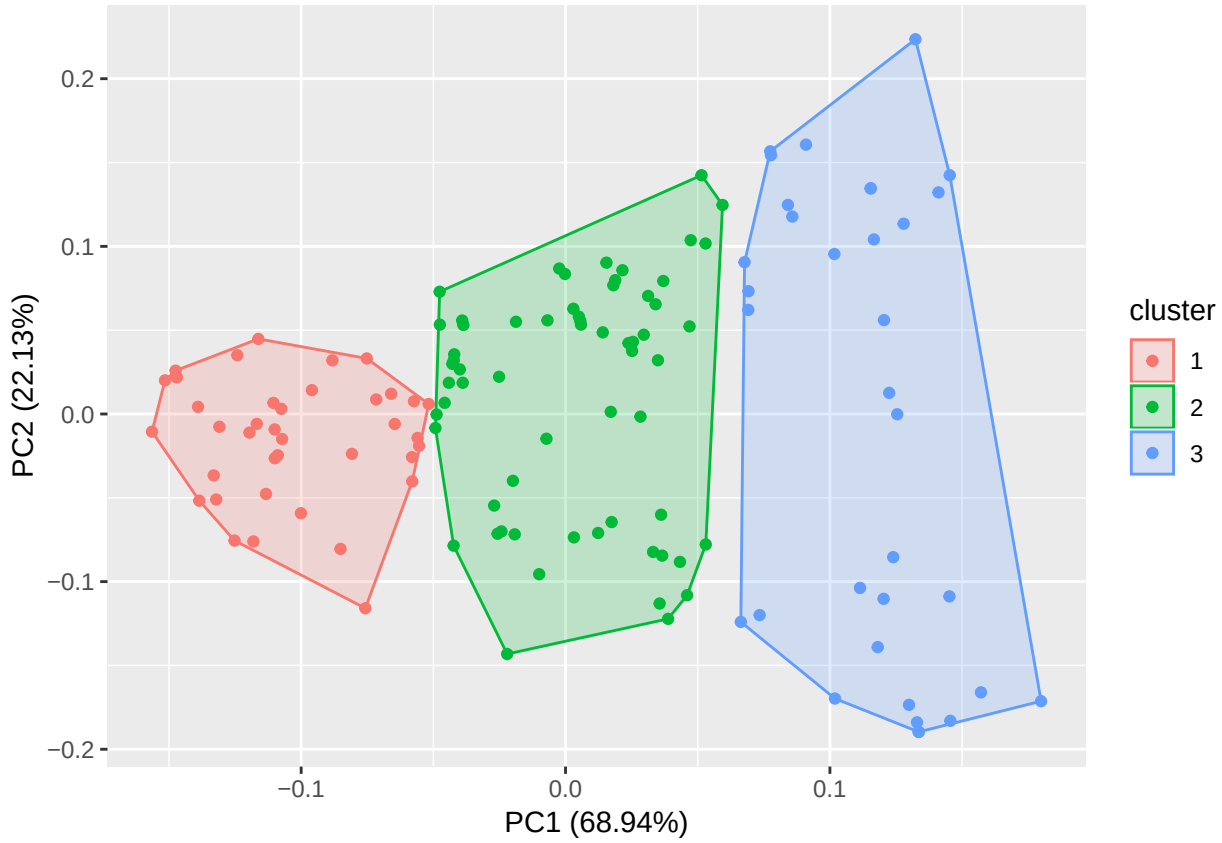


Druhý graf (velmi stručně) udává, do jaké míry spolu prvky ve shluku "sounáleží". Vidíme, že od 2 do přibližně 7 shluků se funkce pohybuje kolem svého maxima.



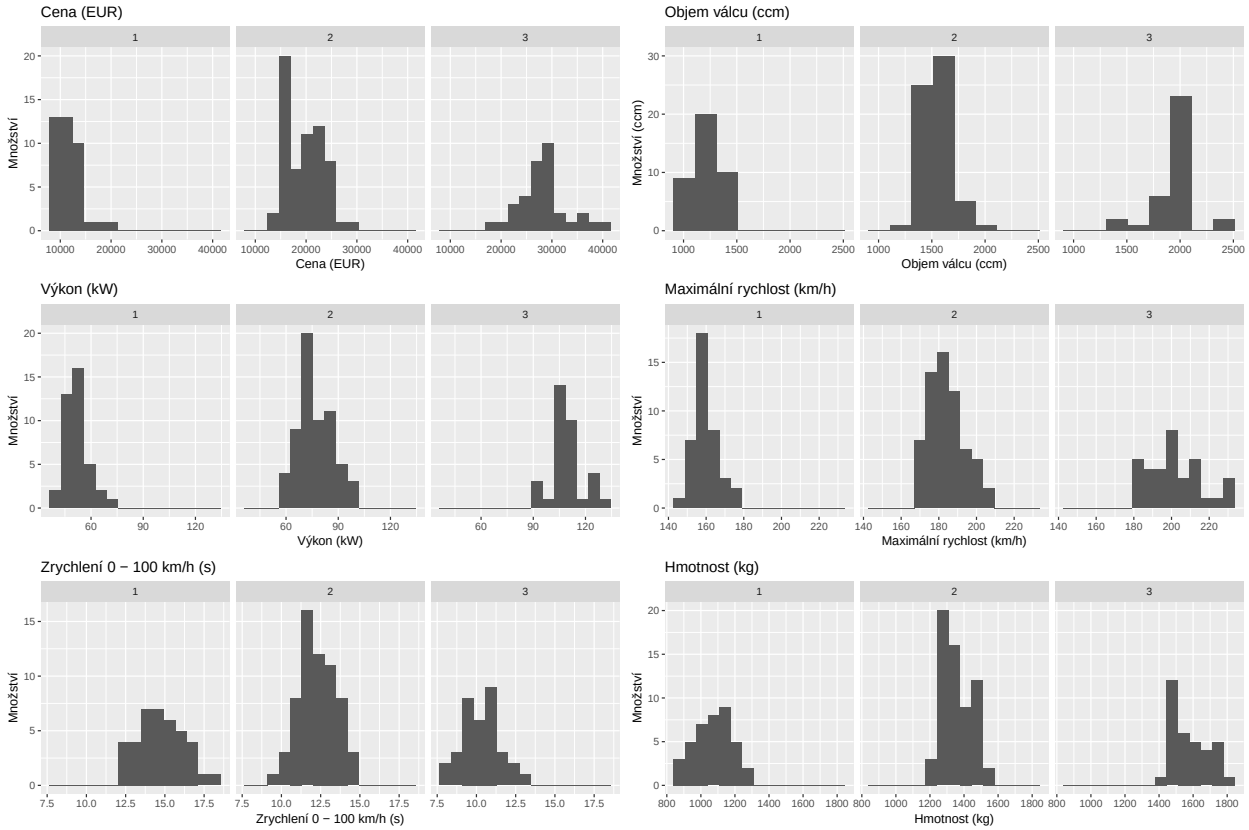
Shluková analýza

Uvažujme tedy shluky tři. Ostatní počty byly rovněž vyzkoušeny, ale zdaleka nejlepší výsledky jsme dostali právě se třemi shluky. Jako proměnné, jež pro segmentaci využijeme, zvolme Maximální rychlost, Výšku, Výkon, Spotřeba - kombinovaná, Objem válců. Tyto proměnné byly vybrány na základě grafu dvou hlavních komponent a zátěží všech proměnných. Vybrali jsme proměnné, které - co se vysvětlení rozptylu týče - spolu nejsou značně korelovány (pozitivně i negativně).



Interpretace shluků

Zobrazme si nyní histogramy jednotlivých proměnných podle shluků a pokusme se je interpretovat.



Interpretace shluků

Shluk 1 (*První vůz - cenová nenáročnost, pohyblivost, mobilita*)

- nízká cena
- malý objem válců
- nízký výkon
- nízká rychlost
- vysoké zrychlení
- malá hmotnost

Interpretace: Tyto modely potěší cenovou nenáročností, nicméně je vykoupena nižším výkonem i nižší rychlostí. Auto však má menší hmotnost, a tak je pohyblivější.

Zástupci: Alfa Romeo MiTo, Fiat Panda, Opel Agila

Shluk 2 (*Kombinace obou světů - rychlý, ovladatelný, výkonný*)

- průměrná cena
- střední objem válců
- střední výkon
- střední rychlost
- střední zrychlení
- střední hmotnost

Interpretace: Tyto modely kombinují nejlepší vlastnosti z obou světů - mají vyšší cenu než nejlevnější kategorie, za to však zákazník dostane vyšší výkon a rychlost.

Zástupci: Mercedes B, Opel Astra GTC, Mazda 6 Sport Kombi

Shluk 3 (***Kvalita a síla***)

- vyšší cena
- vysoký objem válců
- vyšší výkon
- nejvyšší rychlost
- menší zrychlení
- vyšší hmotnost

Interpretace: Zdaleka nejražší modely s bezkonkurenčním výkonem, velkou hmotností a rychlostí. Vyšší hmotnost je vykoupena menším zrychlením.

Zástupci: Citroen c5, Nissan X-Trail