ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE
FAKULTA ELEKTROTECHNICKÁ
Katedra kybernetiky

Predikce spotřeby paliva

Milan Poláček

Zadání

Cílem tohoto úkolu je provést průzkumovou analýzu dat a modelovat závislost spotřeby auta na váze.
Data pro tento úkol byla modifikována z originálních dat v StatLib spravována Carnegie Mellon univerzitou. Originální data o autech byla nasbírána v roce 1980. U každého auta byla zjišťována, spotřeba, výkon, hmotnost, a další charakteristiky, viz Tabulka1
Data jsou k dispozici v souboru auta.csv

1	mpg	spotřeba - kolik mil lze ujet na galon paliva
2	cylinders	Počet válců
3	displacement	Velikost motoru
4	horsepower	Výkon
5	weight	Váha v librách
6	acceleration	Zrychlení
7	modelyear	Rok výroby
8	origin	Indikátor země původu
9	carname	Jméno auta

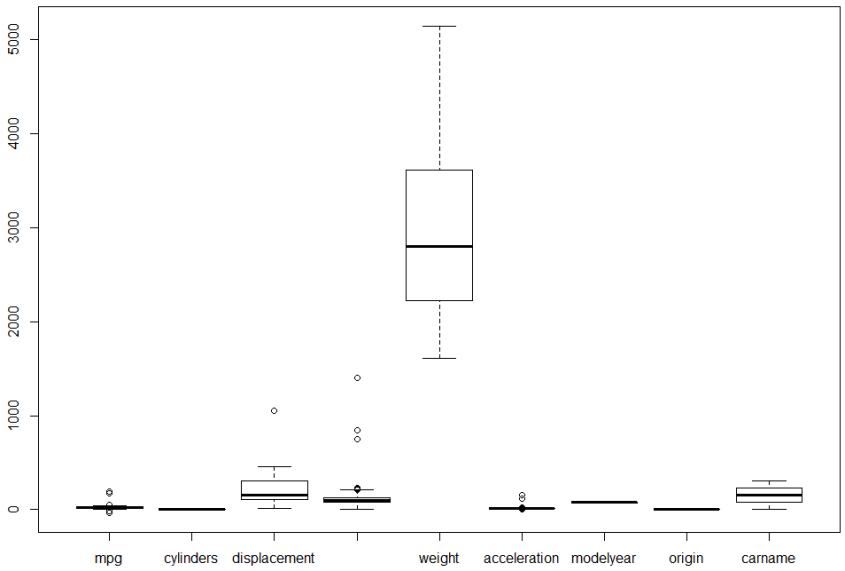
Tabulka 1 Parametry přiložených dat

Požadované kroky analýzy:

1. Kolik máte k dispozici dat (kolik aut, jaké příznaky)? (0.5 bodu)
2. Obsahují data nějaké chybějící hodnoty? (0.5 bodu)
3. Jsou v datech nějaká odlehlá pozorování? Pokud ano, jak se s nimi vypořádáte?(1bod)
4. Vizualizujte vybrané příznaky, vztahy mezi příznaky vzhledem k ostatním bodům úkolu. (2 body)
5. Modelujte závislost spotřeby auta na váze auta a formálně ji zapište. Je statisticky významná? Výsledek slovně interpretujte. (4 body)
6. Existuje i závislost mezi spotřebou a jinými příznaky? (2 body)

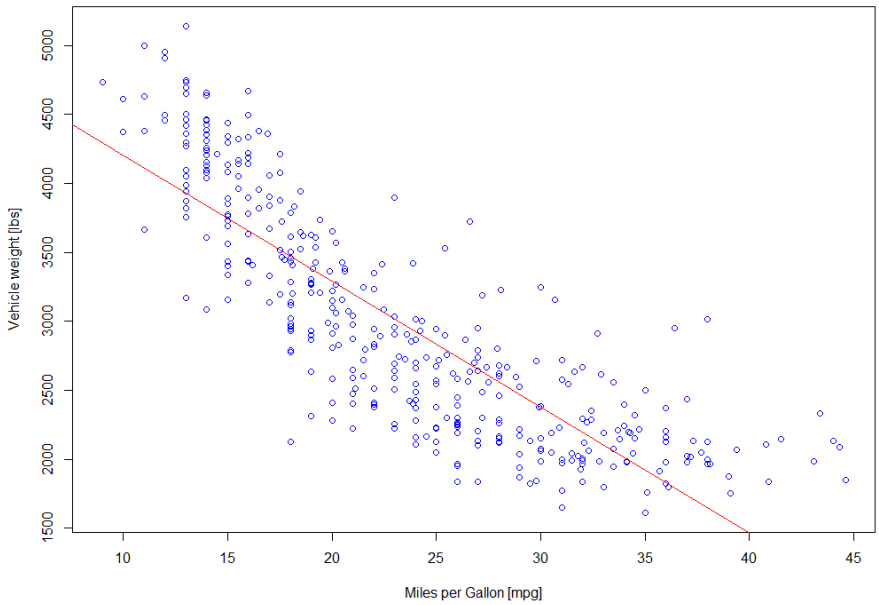
Řešení

K zadání úlohy byla přiložena data v nespecifickém formátu s různými odchylkami ve značení nenaměřených příznaků a chybně zapsaných příznaků.
Jak lze z Tabulka 1 vyčíst máme 9 parametrů (příznaků), které byly měřeny u vozidel. Data obsahovala 398 instancí (data vozidel). Před další analýzou jsem data opravil tak, aby chybně zapsané příznaky, ať už nenaměřené nebo chybně zapsané, měly jednotné značení (NA) a data tak byla připravena pro další analýzu v jazyce R.
Po úpravě značení chybějících příznaků jsem došel k tomu, že 7 instancím (vozidlům) chybí data k velikosti motoru (displacement), 7 instancím chybí parametry o výkonu (horsepower) a 5 instancí nemá data o akceleraci (acceleration). Množiny dat těchto instancí se nepřekrývají.
V datech lze nalézt naměřené příznaky, které jsou odlehlé A jež se vyskytují u 4 příznaků (mpg, displacement, acceleration, horsepower). Pro jejich nalezení jsem využil funkci boxplot viz Graf 1. Tato data lze zanedbat v případě, že počet vyloučených měřených vozidel (instancí) ze statistiky je mnohem menší než celkový počet vozidel. Při zanedbání těchto dat v jednotlivých příznacích nedojde k chybě větší než 4%, což lze považovat za dostačující. Např. v biomedicíně se testy s chybou 5% považují za dostatečně přesné.



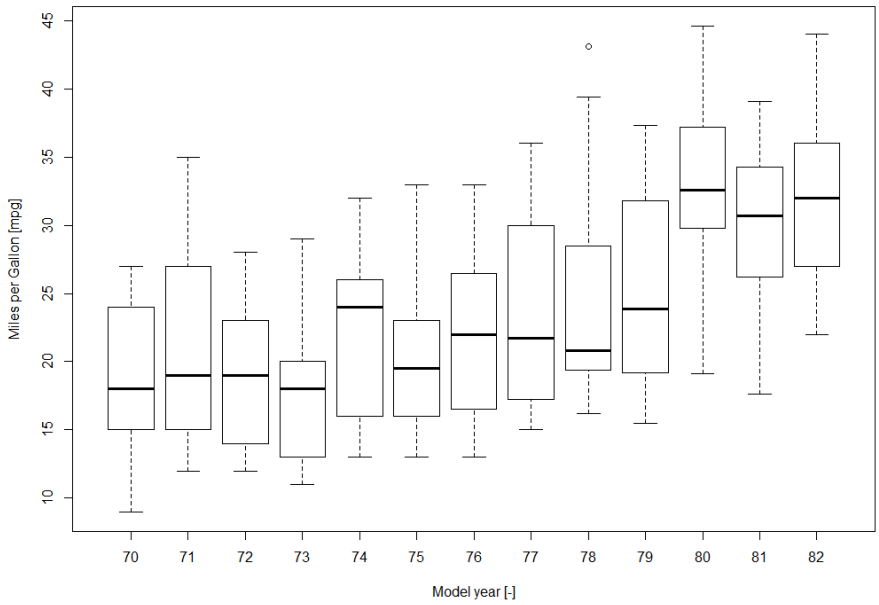
Graf 1 Příznaky a jejich odlehlé hodnoty

Jak je vidět v Graf 2 při vyloučení odlehlých dat lze využít funkce *lm* z jazyka R pro vyjádření závislosti spotřeby na hmotnosti vozidla.
Výsledkem je tedy rovnice přímky (y=k*x+q). Kdy pomocí *coef* jsem zjistil tzv. směrnici přímky (k), která je přibližně -91,3 a koeficient q (určující posun přímky na ose y resp. ose hmotnosti vozidla) přibližně 5116.
Pomocí funkce *anova* z vypočteného modelu pomocí funkce *lm* jsem zjistil, že výsledek je statisticky významný, jak lze z Graf 2 pozorovat, že se snižující se hmotností vozidla se zvyšuje dojezd na jeden galon paliva.

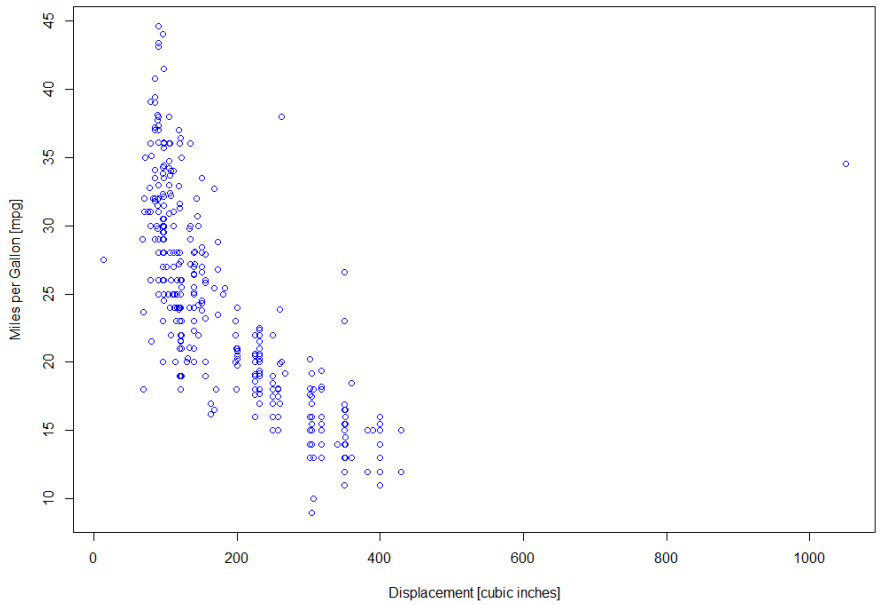


Graf 2 Závislost spotřeby na hmotnosti vybraných vozidel

Dále můžeme například pozorovat závislost spotřeby (resp. dojezdu na galon paliva) na roku výroby vozidla (viz Graf 3) nebo na velikosti motoru (viz Graf 4). Závislost se spotřebou a jinými příznaky tedy lze nalézt.



Graf 3 Závislost spotřeby na roku výroby vozidel



Graf 4 Závislost spotřeby na velikosti motoru vybraných vozidel

Závěr

V tomto úkolu jsem po patřičných úpravách, které jsem popsal výše, analyzoval data naměřená na jednotlivých modelech aut. Následně jsem pro splnění výše uvedených úkolů provedl vizualizaci několika závislostí.
Závislost spotřeby jsem vizualizoval u Graf 2, Graf 3 a Graf 4 bez odlehlých dat příznaku mpg, tak aby byli závislosti okem patrné. Pro názornost odlehlých hodnot jsem v Graf 3 a Graf 4 nechal odlehlé příznaky roku výroby (model year) resp. velikosti motoru (displacement). Bohužel jsem ve své práci mohl zobrazit jen několik zjednodušených grafů, vzhledem k tomu, že jsem v této semestrální práci byl omezen počtem stran.