VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky

Domácí úkoly 1–3

Základy statistiky

Rok: 2021

Cvičící: Ing. Martina Litschmannová, Ph.D.

Autor: Jan Novák (NOV001)

Obsah

1 Domácí úkol 1 3
1.1 Analýza sbírky aut dle typu karoserie 3
2 Domácí úkol 2 4
2.1 Analýza výkonu auta dle typu paliva 4
2.2 Shrnutí výsledků 6
3 Domácí úkol 3 7
3.1 Závislost spotřeby ve městě na výkonu auta 7
3.2 Závislost typu paliva na značce auta 8

1 Domácí úkol 1

Představte si, že data obsahují informace o Vaší soukromé sbírce aut. Jako její hrdý majitel, případně hrdá majitelka, chcete sbírku podrobit drobné explorační analýze.

Analyzujte strukturu vaší sbírky z hlediska typu karoserie auta.

- Součástí popisu musí být tabulka četnosti a alespoň jeden vhodně zvolený grafický výstup.
- Povinnou součástí dokumentu je slovní popis vašeho zjištění, tj. vlastními slovy popište, co jste analýzou zjistili. Tento slovní popis by měl být v rozsahu zhruba 100 slov.

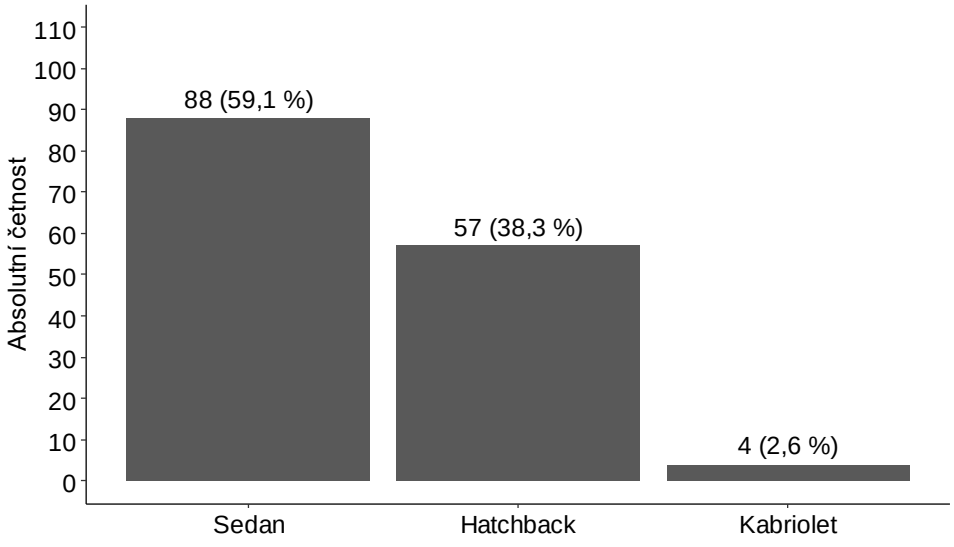
1.1 Analýza sbírky aut dle typu karoserie

Dle zadání bude v této části sbírka aut analyzována z hlediska typu karoserie aut, což je typem tzv. nominální (kategoriální) proměnná. Auta ve sbírce mají jeden z následujících typů karoserie – sedan, hatchback, kabriolet. Pro analýzu jsem zvolil tabulku četnosti s absolutními četnostmi a relativními četnostmi vyjádřenými v procentech doplněnou o sloupcový graf.

V Tab. 1 a na Obr. 1 lze vidět, že nejčetnějším typem karoserie je sedan, který mělo celkem 88 aut ze 149, což činí 59,1 %. Naopak nejméně zastoupeným typem karoserie ve sbírce aut je typ kabriolet, který měla pouze 4 auta ze 149 aut ve sbírce, což činí 2,6 %.

Tab. 1: Struktura sbírky aut dle typu karoserie

	Absolutní četnost	**Relativní četnost (%)**
Sedan	88	59,1
Hatchback	57	38,3
Kabriolet	4	2,6
Celkem	**149**	**100,0**



Obr. 1: Struktura sbírky aut dle typu karoserie (sloupcový graf)

2 Domácí úkol 2

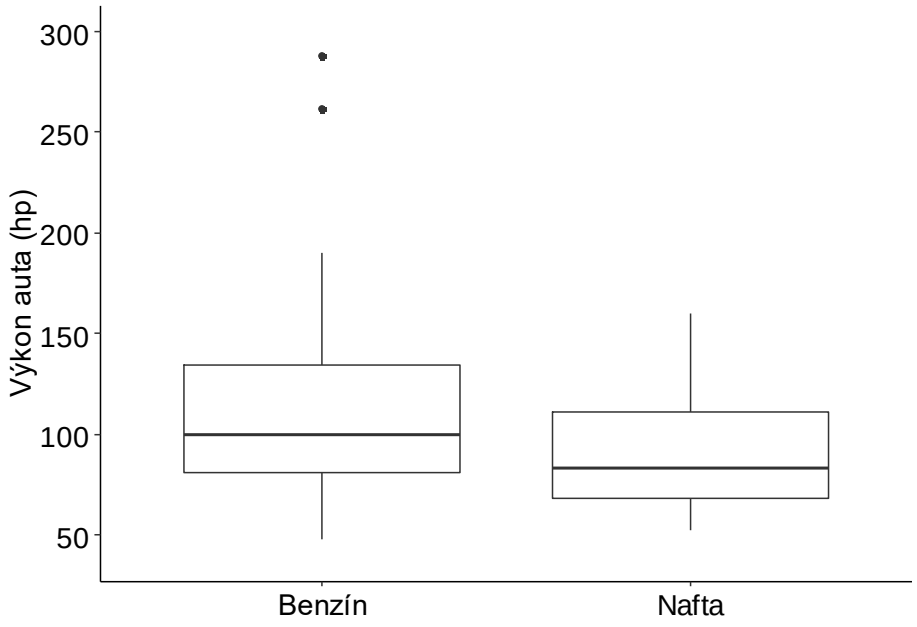
Analyzujte výkon auta s ohledem na jeho typ paliva.

- Součástí analýzy by mělo být rozhodnutí o dalším osudu případných odlehlých pozorováních ve srovnávaných skupinách. Toto rozhodnutí (jejich ponechání nebo odstranění) by mělo být patřičně zdůvodněno.
- Pro srovnání kvantitativní proměnné dle zadané kvalitativní proměnné využijte základní číselné charakteristiky (uveďte je v přehledné tabulce) a vhodnou vizualizaci (vícenásobný krabicový graf, sada histogramů, popř. sada histogramů doplněných o grafy empirické hustoty).
- Povinnou součástí dokumentu je slovní popis vašeho zjištění, tj. vlastními slovy popište, co jste samotnou analýzou zjistili. Tento slovní popis by měl být v rozsahu 200–300 slov (i více).
- Popis by měl čtenáře seznámit s významem uvedených číselných charakteristik v kontextu vaší analýzy a měl by příslušné závěry dát do kontextu s vizualizací problému.

Bonus pro zájemce (možno získat až 2 body navíc): V rámci popisu svých zjištění se pokuste o posouzení, zda je nebo není vhodné modelovat kvantitativní proměnnou normálním rozdělením.

2.1 Analýza výkonu auta dle typu paliva

Nejprve je potřeba zjistit, zda se v původních datech nachází odlehlá pozorování. Za tímto účelem využiji vizualizaci pomocí vícenásobného krabicového grafu. Obr. 2 ukazuje, že se mezi výkony aut, která jezdí na benzín, nachází dvě odlehlá pozorování – je zřejmé, že se jedná o dvě auta, která mají mnohem vyšší výkon než ostatní auta. Stejný závěr poskytuje i identifikace pomocí metody vnitřních hradeb. Pro další analýzu budou tato dvě odlehlá pozorování odstraněna a z původních 91 záznamů aut, která jezdí na benzín, bude využito 89 záznamů. Vzhledem k tomu, že odlehlé hodnoty nenarušují čitelnost vícenásobného krabicového grafu (Obr. 2), nebude už uveden graf prezentující data po jejich odstranění.

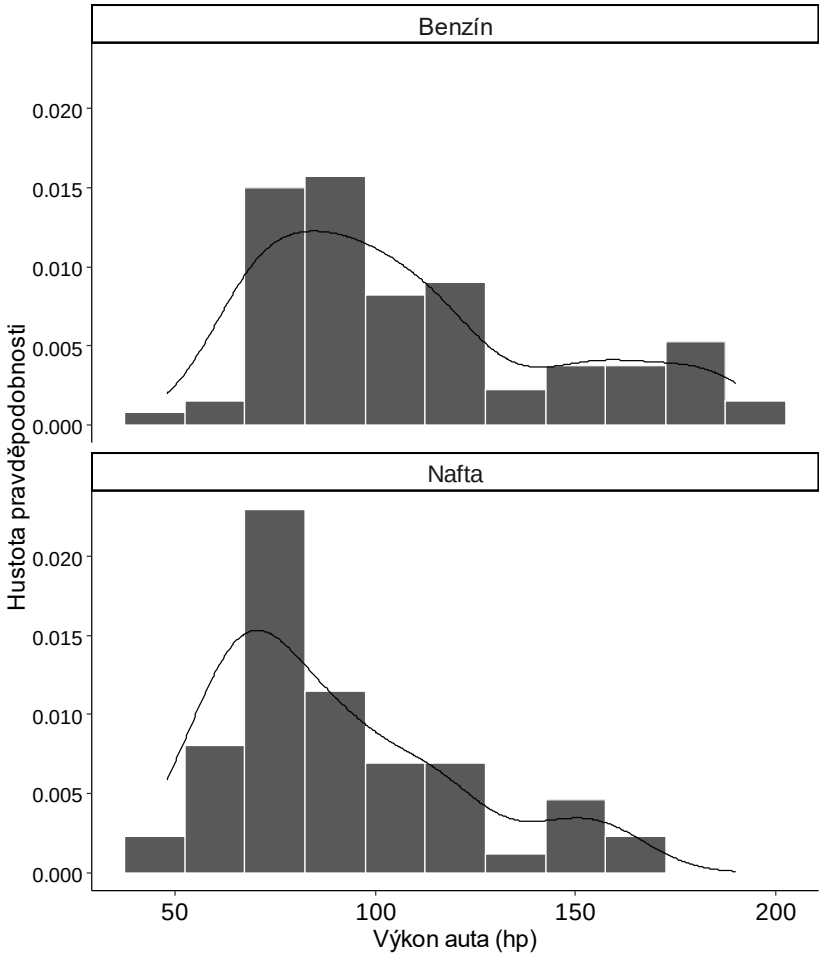


Obr. 2: Vizualizace výkonu auta dle typu paliva (vícenásobný krabicový graf)

Po odstranění identifikovaných odlehlých pozorování lze přistoupit k vlastní analýze dat. Číselné charakteristiky výkonu auta pro obě kategorie určené typem paliva jsou k dispozici v Tab. 2 a kompletní analýzu doplňují histogramy s empirickými hustotami pravděpodobnosti (viz Obr. 3).

Tab. 2: Číselné charakteristiky výkonu (hp) dle typu paliva auta

	Benzín	Nafta
rozsah souboru	89	58
Míry polohy		
minimum	48	52
1. kvartil	78,0	68,0
průměr	108,3	90,4
medián	100,0	83,0
3. kvartil	121,0	110,8
maximum	190	160
Míry variability		
směrodatná odchylka	36,7	29,9
variační koeficient (%)	33,9	33,0
Míry tvaru		
šikmost	0,7	0,9
špičatost	-0,5	-0,2



Obr. 3: Vizualizace výkonu auta dle typu paliva (histogramy s empirickými hustotami pravděpodobnosti)

2.2 Shrnutí výsledků

Výkon aut se zážehovými (benzínovými) motory se pohyboval v rozpětí 48 hp až 190 hp. U poloviny těchto aut se výkon pohybuje v rozmezí 78,0 hp až 121,0 hp. Také lze říct, že polovina aut s benzínovým motorem má výkon alespoň 100,0 hp. Průměrný výkon těchto aut je 108,3 hp a směrodatná odchylka je 36,7 hp. Hodnota variačního koeficientu je 33,9 %, na základě čehož lze považovat výkon aut se zážehovými motory za homogenní. Pokud by auta jezdící na benzín v naší sbírce představovala reprezentativní výběr ze všech aut se zážehovými motory na trhu, pak na základě Čebyševovy nerovnosti by šlo usoudit, že alespoň 75 % aut se zážehovým motorem na trhu má výkon v rozmezí 34,9 hp až 181,7 hp. Analogicky lze interpretovat číselné charakteristiky výkonu aut se vznětovými (naftovými) motory.

Stěžejním bodem této analýzy je srovnání výkonu aut, která jezdí na benzín s těmi, která jezdí na naftu. Na základě vizualizace (Obr. 2, Obr. 3) a číselných charakteristik (Tab. 2) lze usoudit, že auta, která jezdí na benzín mají v průměru vyšší výkon než auta jezdící na naftu. Průměrný výkon aut se zážehovými motory byl 108,3 hp a u vznětových motorů byl průměrný výkon nižší – konkrétně 90,4 hp.

Variabilitu výkonu aut lze považovat v obou kategoriích přibližně za shodnou, což dokládají hodnoty směrodatných odchylek (36,7 hp pro benzínové a 29,9 hp pro naftové motory) i velice podobné hodnoty variačních koeficientů (33,9 % pro benzínové a 33,0 % pro naftové motory). Na tomto místě lze také využít empirického pravidla o poměru výběrových rozptylů srovnávaných souborů. Skutečnost, že poměr výběrových rozptylů je menší než 2 jen dokládá uvedený závěr, že variabilitu výkonu aut lze považovat v obou kategoriích za přibližně shodnou.

Zjištěné míry tvaru uvedené v Tab. 2, tj. šikmost a špičatost, nepoukazují na výrazné odchýlení od normálního rozdělení – obě charakteristiky jsou pro obě kategorie mezi hodnotami -2 a 2. Nicméně, při pohledu na vykreslené histogramy s empirickými hustotami pravděpodobnosti v Obr. 3, je patrné viditelné zešikmení obou histogramů a empirické hustoty pravděpodobnosti zdaleka neodpovídají tvaru Gaussovy křivky. Na základě těchto poznatků se lze spíše přiklánět k tomu, že ani pro jednu kategorii nelze výkon auta modelovat normálním rozdělením.

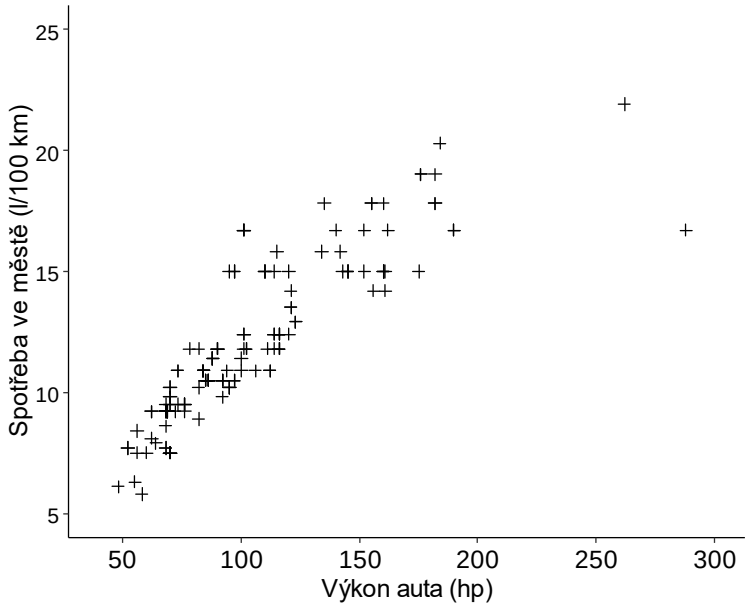
3 Domácí úkol 3

a) Analyzujte závislost výkonu auta na spotřebě ve městě.
b) Analyzujte závislost značky auta a typu paliva.
   - Povinnou součástí dokumentu je slovní popis vašeho zjištění, tj. vlastními slovy popište, jaké nástroje jste pro analýzu závislosti použili a co jste samotnou analýzou zjistili. Tento slovní popis by měl být v rozsahu zhruba 200–300 slov (příp. více).
   - Každou závislost je nutné analyzovat graficky a grafický výstup doplnit o vhodnou tabulku či charakteristiku (kont. tabulka, korelační koeficient, …).

3.1 Závislost spotřeby ve městě na výkonu auta

Spotřeba ve městě a výkon auta jsou kvantitativní proměnné, z toho důvodu pro analýzu jejich závislosti bude využit korelační diagram a vhodný korelační koeficient. V kontextu předchozího domácího úkolu jen upřesním, že opět používám pro tuto analýzu původní data.

Z korelačního diagramu (viz Obr. 4) je jasně viditelná závislost mezi analyzovanými proměnnými. Nicméně dvě pozorování s vysokým výkonem auta se zdají poměrně vzdálená od většiny ostatních. Z toho důvodu určím korelační koeficienty pro původní data i pro data očištěná od těchto dvou odlehlých hodnot (viz Tab. 3).



Obr. 4: Vizualizace závislosti spotřeby auta ve městě na výkonu auta (korelační diagram)

Tab. 3: Korelační koeficienty hodnotící závislost spotřeby auta ve městě na výkonu auta

	Korelační koeficient	
	Pearsonův	**Spearmanův**
Původní data	0,87	0,92
Očištěná data	0,89	0,92

Přítomnost silné pozitivní závislosti naznačují korelační koeficienty v Tab. 3. Díky vysoké hodnotě Pearsonova korelačního koeficientu (0,87 pro původní a 0,89 pro očištěná data) lze dokonce upřesnit, že se jedná o silnou lineární závislost mezi proměnnými (což naznačoval i Obr. 4). Ukázalo se, že odstraněná odlehlá pozorování nebyla tolik vlivná a hodnoty korelačních koeficientů ovlivnila opravdu

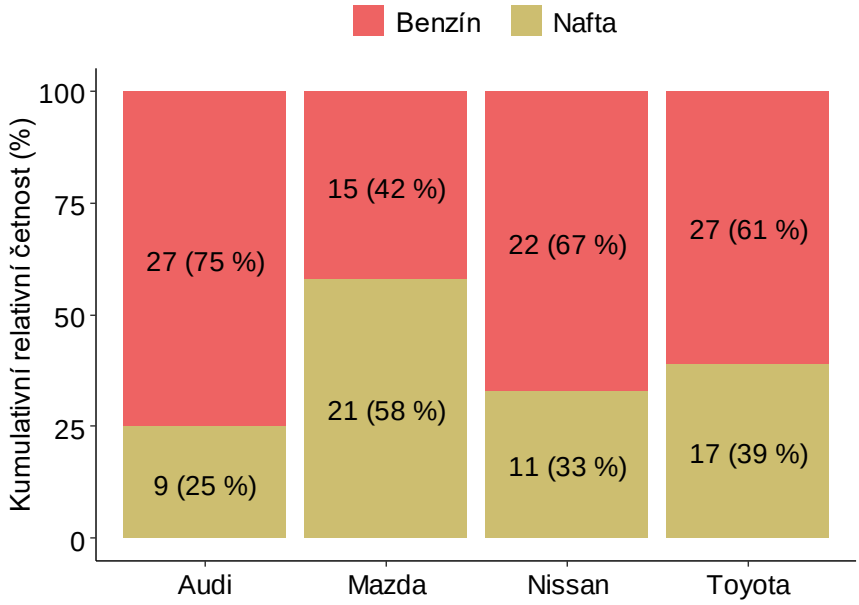
málo. Lze tedy shrnout, že mezi výkonem auta a jeho spotřebou ve městě je silná lineární závislost. S rostoucím výkonem auta lze očekávat rostoucí spotřebu auta ve městě.

3.2 Závislost typu paliva na značce auta

Značka auta a typ paliva jsou kategoriální (nominální) proměnné, z toho důvodu pro analýzu závislosti těchto dvou proměnných využiji kontingenční tabulku a 100% skládaný sloupcový graf.

Tab. 4: Struktura sbírky aut v závislosti na jejich značce a typu paliva (v závorkách jsou uvedeny řádkové rel. četnosti (%))

	Benzín	**Nafta**	**Celkem**
Audi	27 (75)	9 (25)	**36**
Mazda	15 (42)	21 (58)	**36**
Nissan	22 (67)	11 (33)	**33**
Toyota	27 (61)	17 (39)	**44**
Celkem	**91 (61)**	**58 (39)**	**149**



Obr. 5: Struktura sbírky aut v závislosti na jejich značce a typu paliva (100% skládaný sloupcový graf)

Z Tab. 4 a Obr. 5 je patrné, že u 3 ze 4 značek aut převažují auta jezdící na benzín. Jediná značka, u které převažují auta jezdící na naftu je Mazda, která se tím odlišuje od ostatních značek, u kterých převažují benzínové motory. Nejméně aut s naftovým motorem bylo mezi auty značky Audi (25 %). U značky Nissan tvořila auta s naftovým motorem přesně třetinu a u značky Toyota tvořila přibližně dvě pětiny. Z výše uvedeného se zdá, že značku auta a typ paliva lze považovat za závislé statistické znaky, jelikož zastoupení benzínových a naftových motorů napříč značkami není stejné.