4ST426 Regrese, Domácí úkol

Marie Melínová

Uvažujte datový soubor `SaratogaHouses` z balíku `mosaicData`. Vysvětlovanou proměnnou je `price` a ostatní proměnné budou vysvětlující. Datový soubor rozdělte náhodně na data trénovací (cca 80 % všech pozorování) a data testovací (zbývající pozorování).

Cílem analýzy je vytvořit model, pomocí kterého budete sledovat závislost vysvětlované proměnné na proměnných vysvětlujících a pomocí kterého bude možné predikovat cenu nového domu při znalosti hodnot vysvětlujících proměnných.

```
library(mosaicData)
data <- SaratogaHouses

# Rozdělení dat na trénovací a testovací množinu
set.seed(42)
train_index <- sample(1:nrow(data), round(0.8 * nrow(data)), replace = FALSE)

data_train <- data[train_index, ]
# data_test <- data[-train_index, ]
```

Fáze 1

Nalezněte kvantitativní vysvětlující proměnnou, která vykazuje nejvyšší hodnotu indexu determinace v přislušném modelu s proměnnou `price`.

```
tridy_promenych <- ifelse(sapply(data_train[, 2:16], class) == "factor", F, T)
kv_promenne <- names(which(tridy_promenych))

max_indexDeterminace <- 0
max_promenna <- ""

# Vytvoření a porovnání modelů
for (prom in kv_promenne) {

    formula <- paste("price ~", prom)
    model <- lm(formula, data = data_train)
    indexDeterminace <- summary(model)$r.squared

    if (indexDeterminace > max_indexDeterminace) {
        max_indexDeterminace <- indexDeterminace
        max_promenna <- prom
    }
}

c(max_promenna, `Index determinace` = max_indexDeterminace)
```

```
##                       Index determinace
##        "livingArea" "0.507625187778811"
```

Pokud bychom při výběru uvažovali kvadrát výběrového korelačního koeficientu, dospěli bychom ke stejnému výsledku?

- Pokud bychom použili kvadrát výběrového korelačního koeficientu (tj. Pearsonova korelace), pravděpodobně bychom dospěli k **podobnému výsledku, ale ne nutně ke stejnému**.
- Pearsonova korelace měří lineární vztah mezi dvěma proměnnými, zatímco index determinace v lineární regresi měří podíl variability závislé proměnné, který je vysvětlen vysvětlujícími proměnnými.

Fáze 2

Uvažujte regresní model přímky zachycující závislost vysvětlované proměnné na proměnné `bedrooms` a všimněte si znaménka odhadu regresního parametru.

```
model_jednoduchaR <- lm(price ~ bedrooms, data = data_train)
coef(model_jednoduchaR)
```

```
## (Intercept)    bedrooms
##    59333.73    48111.12
```

Nyní uvažujte model vícenásobné lineární regrese zachycující závislost proměnné `price` na proměnných `livingArea` a `bedrooms`.

```
model_vicenasobnaR <- lm(price ~ livingArea + bedrooms, data = data_train)
coef(model_vicenasobnaR)
```

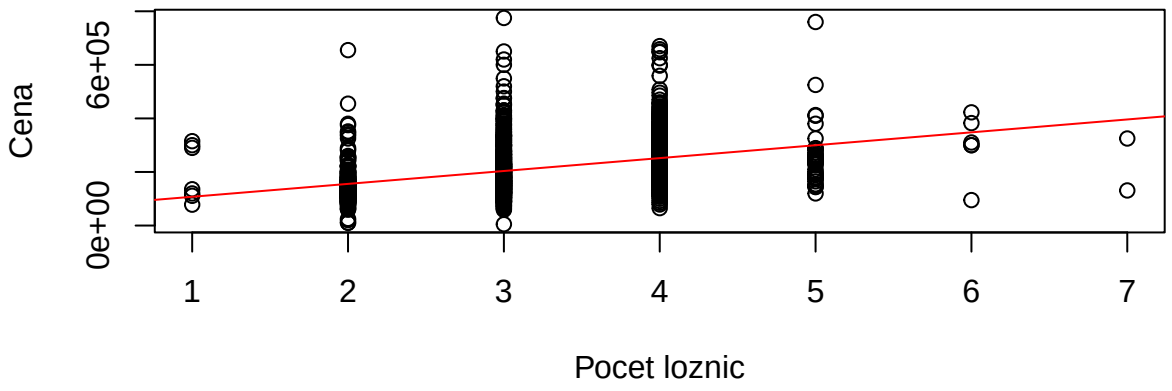
```
## (Intercept)  livingArea    bedrooms
##  36225.5748    122.7533 -12603.6016
```

Změnil se oproti původnímu regresnímu modelu přímky výrazně odhad regresního parametru u proměnné `bedrooms` nebo se dokonce změnilo znaménko tohoto odhadu? Čím si tyto případné změny vysvětlujete?

- Když vysvětlujeme proměnnou `price` pouze proměnnou `bedrooms`, pozitivní regresní koeficient u proměnné `bedrooms` může naznačovat, že čím více ložnic v domě je, tím vyšší je cena. To dává smysl, protože větší počet ložnic obvykle znamená větší dům a mnoho lidí je ochotno platit vyšší cenu za větší bydlení.
- Pokud jsou ložnice a rozloha bytu vzájemně korelované, může se stát, že při zahrnutí obou proměnných do modelu se koeficient u proměnné `bedrooms` stane záporným.

K vysvětlení použijte také vhodně dvou- a třírozměrné grafy.

```
plot(data_train$price ~ data_train$bedrooms, xlab = "Pocet loznic", ylab = "Cena")
abline(model_jednoduchaR, col = "red")
```



K vykreslení 3D grafu bychom použili funkci `plot3d()` z balíčku `rgl`, bohužel při renderování markdownu nelze tato funkce použít a tím pádem ji nemohu zahrnout do řešení domácího úkolu.

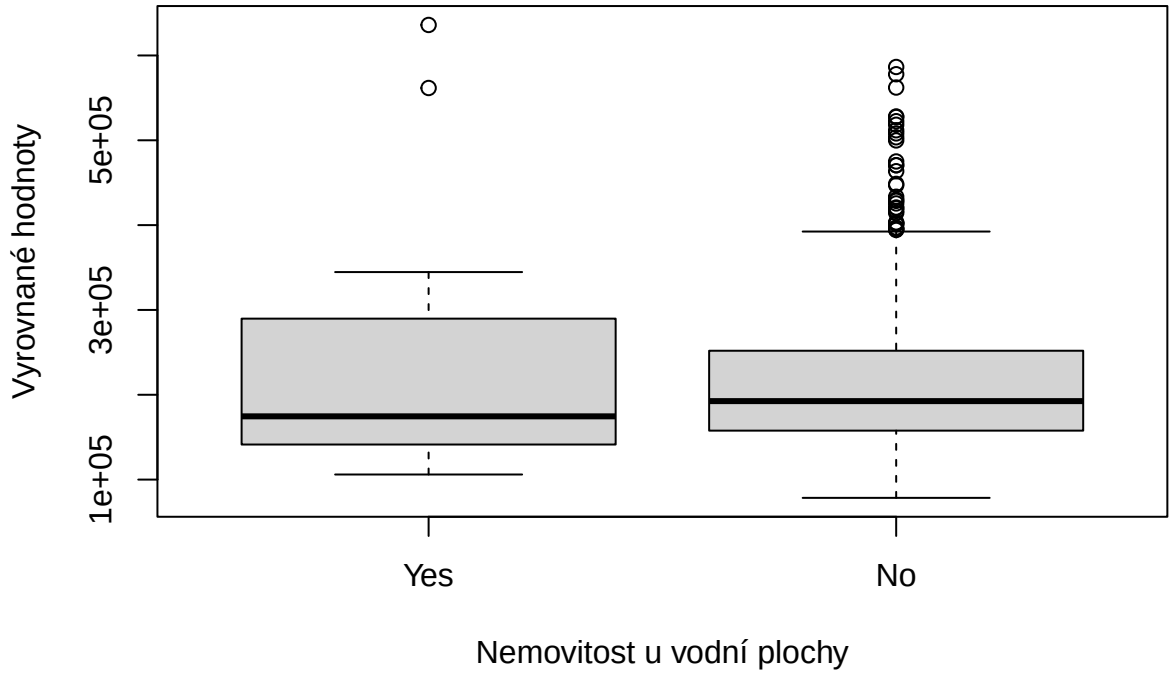
Fáze 3

Uvažujte model vícenásobné lineární regrese zachycující závislost proměnné `price` na všech kvantitativních vysvětlujících proměnných a odhadněte tento model.

```
model_vseKvant <- lm(price ~ lotSize + age + landValue + livingArea + pctCollege +
    bedrooms + fireplaces + bathrooms + rooms, data = data_train)
```

1. Pracujte s vyrovnanými hodnotami, na které vykreslete krabičkové grafy odděleně pro kategorie (Yes a No) kategoriální proměnné `waterfront`. Tato proměnná uvádí, zda se nemovitost nalézá u vodní plochy či nikoli.

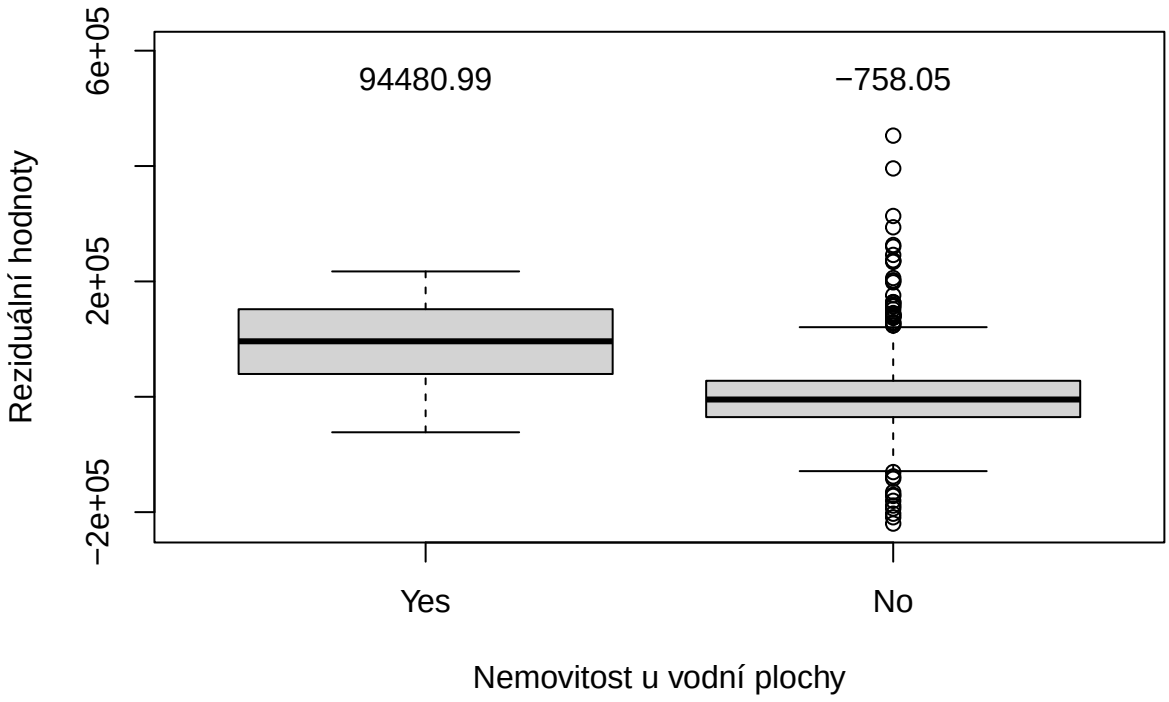
```
plot(fitted(model_vseKvant) ~ interaction(data_train$waterfront), ylab = "Vyrovnané hodnoty",
    xlab = "Nemovitost u vodní plochy")
```



2. Následně pro tento model nalezněte průměrnou hodnotu reziduí obou kategorií zvlášť. Vyneste rovněž krabičkové grafy reziduí.

```
plot(resid(model_vseKvant) ~ interaction(data_train$waterfront), ylab = "Reziduální hodnoty",
    xlab = "Nemovitost u vodní plochy", ylim = c(min(resid(model_vseKvant)), 6e+05))

text(1, 550000, round(mean(resid(model_vseKvant)[which(data_train$waterfront == "Yes")]),
    2))
text(2, 550000, round(mean(resid(model_vseKvant)[which(data_train$waterfront == "No")]),
    2))
```



3. Interpretujte výsledky a vysvětlete z nich např. některé skutečnosti vztahu proměnné `price` a `waterfront`.

- Reziduální hodnoty nemovitostí, které se nachází u vodní plochy, jsou o mnoho vyšší, než reziduální hodnoty nemovitostí, které se u vodní plochy nenachází.
- Do modelu bychom tedy tuto proměnnou mohli zahrnout, jelikož proměnná `price` se jeví jako závislá i na proměnné `waterfront`.

Fáze 4

Uvažujte model vícenásobné lineární regrese zachycující závislost proměnné `price` na všech kvantitativních a také kategoriálních vysvětlujících proměnných. Zatím však neuvažujte interakce a nelinearity. Kategoriální proměnné reprezentujte v modelu s využitím dummy proměnných. Odhadněte tento model a interpretujte odhad regesního parametru u dummy proměnné reprezentující kategoriální proměnnou `centralAir`, tj. přítomnost klimatizace. Je znaménko a hodnota tohoto odhadu v souladu s původní představou?

```
# Kategoriální proměnné jsou kódované jako faktory, takže do regresního modelu
# automaticky vstupují jako dummy proměnné

model_vse <- lm(price ~ ., data = data_train)
round(coef(model_vse), 3)
```

```
##            (Intercept)                lotSize                    age
##              91145.032              8471.303               -161.825
##              landValue             livingArea             pctCollege
##                  0.915                67.096                -92.049
##               bedrooms             fireplaces              bathrooms
##              -7180.005              -720.765              22667.978
##                  rooms heatinghot water/steam        heatingelectric
##               3301.101             -8035.716              -7573.326
##           fuelelectric               fueloil sewerpublic/commercial
##              -3654.033             -4061.107              -1592.427
##              sewernone          waterfrontNo      newConstructionNo
##             -21486.338            -95032.408              40993.077
##           centralAirNo
##              -9983.026
```

- Za předpokladu, že bychom měli dva identické domy, ale jeden s klimatizací a druhý bez, tak dům, který klimatizaci nemá, bude v průměru o 9983.03 $ levnější, než dům s klimatizací.
- Znaménko je v souladu s původní představou.

Uvažujte pouze ta pozorování, pro něž přítomnost klimatizace nabývá hodnoty `Yes`. S využitím těchto pozorování odhadněte model zachycující závislost proměnné `price` na všech kvantitativních i kategoriálních vysvětlujících proměnných s výjimkou `centralAir`. Následně pracujte pouze s těmi pozorováními, jež odpovídají nemovitostem bez klimatizace. Pro tato pozorování odhadněte model při zahrnutí stejných proměnných jako v předchozím případě. Porovnejte odhady a vysvětlete, jaký vliv má přítomnost klimatizace na cenu domu. Liší se původní představa a výsledek?

```
library(car)
```

```
## Loading required package: carData
```

```
model_sKlima <- lm(price ~ ., data = data_train[data_train$centralAir == "Yes", 1:15])
model_bezKlima <- lm(price ~ ., data = data_train[data_train$centralAir == "No",
    1:15])

compareCoefs(model_sKlima, model_bezKlima, model_vse)
```

```
## Calls:
## 1: lm(formula = price ~ ., data = data_train[data_train$centralAir == "Yes",
##    1:15])
## 2: lm(formula = price ~ ., data = data_train[data_train$centralAir == "No",
```

```
##   1:15])
## 3: lm(formula = price ~ ., data = data_train)
##
##                         Model 1 Model 2 Model 3
## (Intercept)               91496  107777   91145
## SE                        42285   25295   22351
##
## lotSize                    6722    9057    8471
## SE                         4664    3018    2589
##
## age                       -652.1  -128.5  -161.8
## SE                         239.9    63.5    65.2
##
## landValue                0.7973  0.9506  0.9154
## SE                       0.0804  0.0701  0.0512
##
## livingArea                84.81   53.91   67.10
## SE                         9.80    5.75    5.07
##
## pctCollege               -307.2    58.9   -92.0
## SE                        396.7   171.3   164.2
##
## bedrooms                 -12818   -1930   -7180
## SE                         5402    3208    2831
##
## fireplaces                -6355    3853    -721
## SE                         6057    3744    3250
##
## bathrooms                 25518   17812   22668
## SE                         6704    4278    3662
##
## rooms                      3918    2425    3301
## SE                         1945    1198    1048
##
## heatinghot water/steam    -5840   -5872   -8036
## SE                        14680    4590    4656
##
## heatingelectric          -13921   -6390   -7573
## SE                        19386   23291   13582
##
## fuelelectric              -9638   -1550   -3654
## SE                        18169   23219   13390
##
## fueloil                   17054   -7791   -4061
## SE                        13587    5719    5540
##
## sewerpublic/commercial   -11844    2387   -1592
## SE                         8352    4440    4056
##
## sewernone                -26048  -18573  -21486
## SE                        33254   21396   18393
##
## waterfrontNo             -98456 -101329  -95032
## SE                        34077   20007   17799
```

```
##
## newConstructionNo        57595   23304   40993
## SE                       12396   11289    8130
##
## centralAirNo                             -9983
## SE                                        3791
##
```

Co lze konstatovat o celkové analýze závislosti `price` na všech vysvětlujících proměnných v porovnání s analýzou v podsouborech rozdělených podle přítomnosti klimatizace?

- Na základě rychlé (a pouze vizuální) analýzy rozdílnosti jednotlivých koeficientů si můžeme všimnout několika zajímavých věcí, jako např.:
  - V domech, kde se nachází klimatizace, se negativně cení přítomnost krbu, zatímco v domech, kde klimatizace není, se přítomnost krbu cení spíše pozitivně,
  - V domech, kde se nachází klimatizace, se přibližně 1.6-krát více cení vyšší počet pokojů, než v domech, kde klimatizace není,
  - atd.

Fáze 5

Nejprve slovně na základě věcné úvahy i intuice zdůvodněte, jaké vysvětlující proměnné by měly být určující z hlediska ceny nemovitosti. Nevyužívejte zde data ani předchozí analýzy. Diskutujte, zda by vysvětlující proměnné měly vstupovat do modelu lineárně anebo nikoli a zda budou přítomny interakce. Na tomto základě navrhněte matematickou rovnici a zapiště model vícenásobné lineární regrese při zohlednění nelinearit vysvětlujících proměnných a interakcí. V navrženém modelu musí být zařazena alespoň jedna vysvětlující proměnná, která vstupuje nelineárně ve formě polynomu či splinu a alespoň jedna interakce.

1. **Rozloha domu (`livingArea`)**: Většinou platí, že větší domy mají vyšší cenu, protože nabízejí více prostoru pro bydlení a jsou obvykle žádanější (tato proměnná bude do modelu vstupovat ve formě polynomu - od určité rozlohy už další jednotka nemusí být tak ceněná).
2. **Počet ložnic (`bedrooms`)**: U domů s více ložnicemi a koupelnami je obvykle vyžádána vyšší cena, protože poskytují vyšší komfort a větší flexibilitu pro rodiny nebo jednotlivce.
3. **Stáří domu (`age`)**: Starší domy mohou mít určitý historický nebo estetický význam, což může zvyšovat jejich hodnotu. Na druhou stranu, nově postavené domy mohou nabízet moderní vybavení a mohou být více energeticky efektivní.
4. **Hodnota pozemku (`landValue`)**: Hodnota pozemku může odrážet atraktivitu a prestiž dané lokality. Výhodná lokalita může odrážet vyšší cenu nemovitosti kvůli svým vlastnosem, jako je blízkost k veřejné dopravě, školám, parkům nebo obchodům.

Dále do modelu zařadíme dvě interakce:

- **Interakce mezi `livingArea` a `bedrooms`**: Tato interakce by mohla zohledňovat skutečnost, že větší domy mohou mít tendenci mít více ložnic - to může zachytit situaci, kdy se cena domu zvyšuje tím víc, čím větší je plocha a čím víc koupelen se v domě nachází.
- **Interakce mezi `livingArea` a `age`**: Tato interakce může podchytit situaci, kdy vztah mezi cenou domu a jeho stářím se mění v závislosti na jeho velikosti.
  - Například, pokud má starší dům velkou plochu, může být jeho cena relativně vyšší než u starších domů menších rozměrů, protože velké starší domy mohou mít např. historickou hodnotu.

$$price = \beta_0 + \beta_1 \cdot livingArea + \beta_2 \cdot livingArea^2 + \beta_3 \cdot bedrooms + \beta_4 \cdot age + \beta_5 \cdot landValue +$$
$$\beta_6 \cdot livingArea \times bedrooms + \beta_7 \cdot livingArea \times age$$

Navržený model odhadněte, interpretujte vlilv proměnných, jež vstupovaly do modelu nelineárně a interpretujte interakce. Předpokládejte splnění všech předpokladů a vhodnými testy hypotéz posuďte, zda bylo užití nelinearit a interakcí třeba.

```
model <- lm(price ~ poly(livingArea, degree = 2) + bedrooms + age + landValue + livingArea:bedrooms +
    livingArea:age, data = data_train)

summary(model)
```

```
##
## Call:
## lm(formula = price ~ poly(livingArea, degree = 2) + bedrooms +
##     age + landValue + livingArea:bedrooms + livingArea:age, data = data_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -240485  -35175   -4665   27343  465528
##
## Coefficients:
```

```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    1.998e+05  9.011e+03  22.171  < 2e-16 ***
## poly(livingArea, degree = 2)1  2.557e+06  3.238e+05   7.896 5.86e-15 ***
## poly(livingArea, degree = 2)2  1.032e+05  8.362e+04   1.234   0.2176
## bedrooms                       4.356e+03  7.567e+03   0.576   0.5650
## age                            1.181e+02  1.522e+02   0.776   0.4380
## landValue                      9.178e-01  5.099e-02  17.998  < 2e-16 ***
## bedrooms:livingArea           -4.691e+00  4.039e+00  -1.162   0.2456
## age:livingArea                -2.271e-01  8.001e-02  -2.838   0.0046 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59530 on 1374 degrees of freedom
## Multiple R-squared:  0.6194, Adjusted R-squared:  0.6174
## F-statistic: 319.4 on 7 and 1374 DF,  p-value: < 2.2e-16
```

Za předpokladu, že všechny ostatní proměnné zůstávají konstantní,

- koeficient β_1 nelze přesně interpretovat jako v lineární regresi, jelikož nelze vymezit situaci, kdy se změní pouze jeden koeficient bez změny druhého,
- koeficient β_2 představuje příspěvěk kvadratického členu na cenu nemovitosti - je relativně menší než β_1, což naznačuje, že nárůst ceny nemovitosti v důsledku zvětšení plochy domu se snižuje s rostoucí plochou,
- koeficient β_6 udává, jak se změní efekt plochy na cenu nemovitosti s každou jednotkovou změnou počtu ložnic - koeficient je negativní, což naznačuje, že nárůst plochy domu má menší přínos k ceně nemovitosti v domech s větším počtem ložnic, než v domech s nižším počtem ložnic,
- koeficient β_7 udává, jak se změní efekt plochy na cenu nemovitosti s každým rokem stárnutí domu - koeficient je opět negativní, což znamená, že hodnota přidaná každou další jednotkou plochy k ceně nemovitosti klesá s rostoucím stářím domu.

```r
model_2 <- lm(price ~ livingArea + bedrooms + age + landValue, data = data_train)
anova(model, model_2)
```

```
## Analysis of Variance Table
##
## Model 1: price ~ poly(livingArea, degree = 2) + bedrooms + age + landValue +
##     livingArea:bedrooms + livingArea:age
## Model 2: price ~ livingArea + bedrooms + age + landValue
##   Res.Df        RSS Df   Sum of Sq      F  Pr(>F)
## 1   1374 4.8700e+12
## 2   1377 4.9138e+12 -3 -4.3828e+10 4.1218 0.00638 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na klasické hladině významnosti nezamítáme nulovou hypotézu ve tvaru $H_0 : \beta_2 = \beta_6 = \beta_7 = 0$. Mohli bychom tedy říct, že použití nelinearit a interakcí nebylo potřeba.

Fáze 6

Vyjděte z modelu, který jste navrhli v předchozí fázi a posuďte, které vysvětlující proměnné jsou přínosné pro zjištění ceny domu a které naopak přínosné nejsou. Cílem není posuzovat statistickou významnost odhadnutých parametrů pomocí testů hypotéz, ale praktickou významnost (např. ekonomický dopad) či významnost ve smyslu rozkladu regresního součtu čtverců, resp. indexu determinace. Shodují se tyto úvahy s výsledky statistické analýzy?

```
a <- anova(model)

prinos_promennych <- as.data.frame(matrix(nrow = 7, ncol = 0))
prinos_promennych$PROM <- c("livingArea", "livingArea^2", "bedrooms", "age", "landValue",
    "livingArea:bedroom", "livingArea:age")
prinos_promennych$PRINOS <- round(a$`Sum Sq`/sum(a$`Sum Sq`) * 100, 2)
# Přínos jednotlivých proměnných v % vysvětlené variability

prinos_promennych
```

```
##                 PROM PRINOS
## 1         livingArea  51.05
## 2       livingArea^2   0.51
## 3           bedrooms   0.40
## 4                age   9.65
## 5          landValue   0.11
## 6 livingArea:bedroom   0.22
## 7     livingArea:age  38.06
```

Největší podíl na vysvětlené variabilitě má proměnná `livingArea` a interakce `livingArea:age`. U proměnné `livingArea` se není čemu divit, už v první fázi jsme zjistili, že proměnnou `price` vysvětluje nejlépe právě tato proměnná.

Další proměnné mají podíl na vysvětlené variabilitě jen velmi nízký, proto se není čemu divit, že nám statistický test v minulé fázi potvrdil, že se nelineární vztahy a interakce nemuseli použít.

Fáze 7

Nalezněte model, pomocí kterého budete co nejpřesněji předpovídat ceny domu pro nová pozorování. Při hledání modelu můžete využít postupy typu regresních polynomů, splinů, interakcí a různých kritérií pro výběr modelu: upravený index determinace, informační kritéria, PRESS statistiku či křížovou validaci. Můžete vyřazovat odlehlá pozorování a využít postupy uvedené v doporučené literatuře, např. hřebenovou regresi či regresi na hlavních komponentech. Výběr modelu provádějte výhradně na trénovacích datech. Následně na testovacích datech odhadněte očekávanou čtvercovou chybu předpovědi ceny nemovitosti pro finální model, jež jste vybrali na základě trénovacích dat.