1. zápočtová úloha z 01RAD 2022/23

Popis úlohy

V tomto úkolu je cílem provést předzpracování datového souboru, jeho vizualizaci a jednoduchou lineární regresní úlohu, kde se budeme zajímat o ceny nemovitostí. Za tímto účelem využijeme datový set *saratosa_hoouses* z knihovny *moderndive* obsahující výběr 1057 domů.

Podmínky a body

Úkol i protokol vypracujte samostatně. Pokud na řešení nějaké úlohy budete přesto s někým spolupracovat, radit se, nezapomeňte to u odpovědi na danou otázku uvést. Tato zápočtová úloha obsahuje 10 otázek po 1 bodu. Celkem za 3 zápočtové úlohy bude možné získat 30 bodů, přičemž pro získání zápočtu je potřeba více jak 20 bodů. Další dodatečné body mohu případně individuálně udělit za řešení mini domácích úkolů z jednotlivých hodin.

Odevzdání

Protokol ve formátu Rmd+pdf, nebo jako Jupyter notebook (idealně odkaz na gitlab s možností spustit v Colabu) nejpozději do 1. 11. 2022.

Předzpracování dat:

```
list_of_packages <- c("tidyverse", "MASS","GGally","moderndive")
missing_packages <- list_of_packages[!(list_of_packages %in% installed.packages()[,"Package"])]
missing_packages
```

```
## character(0)
```

```
if(length(missing_packages)) install.packages(missing_packages)
lapply(list_of_packages, library, character.only = TRUE)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
##
## Attaching package: 'MASS'
##
##
## The following object is masked from 'package:dplyr':
##
##     select
##
##
## Registered S3 method overwritten by 'GGally':
```

```
##   method from
##   +.gg   ggplot2
```

```
## [[1]]
##  [1] "forcats"   "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"
##  [7] "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices"
## [13] "utils"     "datasets"  "methods"   "base"
##
## [[2]]
##  [1] "MASS"      "forcats"   "stringr"   "dplyr"     "purrr"     "readr"
##  [7] "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"
## [13] "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[3]]
##  [1] "GGally"    "MASS"      "forcats"   "stringr"   "dplyr"     "purrr"
##  [7] "readr"     "tidyr"     "tibble"    "ggplot2"   "tidyverse" "stats"
## [13] "graphics"  "grDevices" "utils"     "datasets"  "methods"   "base"
##
## [[4]]
##  [1] "moderndive" "GGally"     "MASS"       "forcats"    "stringr"
##  [6] "dplyr"      "purrr"      "readr"      "tidyr"      "tibble"
## [11] "ggplot2"    "tidyverse"  "stats"      "graphics"   "grDevices"
## [16] "utils"      "datasets"   "methods"    "base"
```

```
? saratoga_houses
```

```
## starting httpd help server ... done
```

Data

```
head(saratoga_houses)
```

```
## # A tibble: 6 x 8
##    price living_area bathrooms bedrooms fireplaces lot_size   age fireplace
##    <dbl>       <dbl>     <dbl>    <dbl>      <dbl>    <dbl> <dbl> <lgl>
## 1 142212        1982       1          3          0     2      133 FALSE
## 2 134865        1676       1.5        3          1     0.38    14 TRUE
## 3 118007        1694       2          3          1     0.96    15 TRUE
## 4 138297        1800       1          2          2     0.48    49 TRUE
## 5 129470        2088       1          3          1     1.84    29 TRUE
## 6 206512        1456       2          3          0     0.98    10 FALSE
```

Otázka 01

Zjistěte, zdali data neobsahují chybějící hodnoty *NA*. Pokud ano, tak rozhodněte zdali můžete příslušná pozorování z dat odstranit a proč. Které proměnné jsou kvantitativní a které kvalitativní? Jeli možno některé zařadit do obou skupin, pro kterou byste se rozhodli? Které proměnné je možné pužít jako faktorové ordinální a jaké jako faktorové nominální a proč? Spočtěte základní statistiky pro jednotlivé proměnné.

Řešení 01:

Otázka 02

Chceme koupit nemovitost v zahraničí a průzkumem trhu jsme obdřeli předchozí data set *saratoga_houses*. Jelikož ale máme přesnější požadavky a nerozumíme imperiálním jednotkám, potřebujeme data upravit:

- Převeďte cenu nemovitostí z dolarů na koruny v tisících a plochu pozemku a obytnou plochu z akrů a čtverečích stop na m^2. (chceck describtion by *? saratoga_houses*)

- Vyberte jen nemovitosti starší 10 let a mladší 50 let, jejichž cena je menší než 7500000 Kč, a plocha pozemku je mezi 500 a 5000 m^2.
- Počet koupelen a počet pokojů převeďte na faktorové proměnné o 3 úrovních.

Dále pracujte jen s takto omezeným datasetem a s proměnnýma *cena*, *plocha_obytna*, *plocha_pozemku* *pocet_pokoju*,*stari_domu*, *pocet_koupelen*, *krb*.

Řešení 02:

Otázka 03

- Porovnejte průměry cen nemovitostí s krbem a bez krbu a otestujte, zdali na hladině významnosti $\alpha = 0.01$ je průměrná cena nemovitostí s krbem větší než průměrná cena nemovitostí bez krbu.

Řešení 03:

Vizualizace dat

Otázka 04

- Vykreslete scatterploty pro všechny numerické proměnné, kde bude barevně rozlišeno, zdali se jedná o nemovitost s krbem, nebo bez krbu.
- Pro proměnné *pocet_pokoju* a *pocet_pater* a *sklep* vykreslete krabicové diagramy (nebo violin ploty), kde odezvou bude *cena*.
- Pro proměnnou *cena* vykreslete histogram spolu s jádrovým odhadem hustoty.

Otázka 05

Pro kombinace faktorizovaných proměnných *pocet_pokoju*, *pocet_koupelen* vykreslete cenu nemovitosti, aby bylo na obrázku vidět, jestli se v průměru liší ceny nemovitostí majících více pokojů, nebo více koupelen a zdali jsou zastoupeny všechny kombiance všech úrovních pro dvě zmíněné faktorové proměnné.

Otázka 06

Pro nemovitosti s dvěma ložnicema vykreslete závislost ceny na obytné ploše nemovitosti, kde jednotlivé události označíte barvou podle toho zdali mají krb a velikost bodů v grafu bude odpovídat počtu koupelen (pro tuto úlohu je lepší vzít počet koupelen jako numerickou proměnnou).

Dále pracujte jen s nemovitostmi se dvěma ložnicemi.

Jednoduchý lineární model

Otázka 07

Sestavte jednoduchý regresní model (s i bez interceptu), kde vysvětlovaná proměnná bude cena nemovitosti a vysvětlující obytná plocha. Spočtěte pro oba modely R^2 a F statistiky, co nám o modelech říkají. Vyberte jeden z nich a zdůvodněte proč ho preferujete.

Na základě zvoleného modelu odpovězte, zdali cena nemovitosti závisí na obytné ploše a pokud ano, o kolik se změní očekávaná cena pro nemovitost s obytnou plochou zvětšenou o $20m^2$?

Otázka 08

Sestavte jednoduchý linární model jako v předchozí otázce pro nemovitosti s krbem a bez krbu. Jaký model vykazuje silnější linearní vztah mezi cenou a obytnou plochou? O kolik cena s rostoucí obytnou plochou pro nemovitosti s krbem roste rychleji než pro nemovitosti bez krbu?

Spočtěte 95% konfidenční intervaly pro regresní koeficienty popisující sklon regresní přímky v obou modelech a zjistěte, zdali se protínají. Co z toho můžeme vyvozovat?

Na základě těchto modelů zjistěte o kolik procent bude mít průměrná nemovitost s krbem a obytnou plochou $160m^2$ vyšší očekávanou cenu než průměrná nemovitost o stejné obytné ploše, ale bez krbu.

Otázka 9

Vykreslete scatterplot obytné plochy a ceny nemovitostí. Do tohoto grafu vykreslete regresní přímky vybraných modelů pro nemovitosti s krbem a bez něho, jednoltivé body i regresní přímky označte barvou podle toho k jaké skupině přísluší.

Sestrojte 90% konfidenční intervaly okolo očekávaných cen pro jednoltivé skupiny a na jejich základě rozhodněte, zdali a jak se očekávané ceny budou lišit pro nemovitosti s obytnou plochou menší než $120m^2$. Je to porovnávání správné? Zdůvoněte.

Otázka 10

Vykreslete histogramy pro rezidua modelů z předchozí otázky. Proložte je hustotou normálního rozdělení s nulovou střední hodnotou a rozptylem odpovídajícím $\hat{\sigma}^2$ z jednotlivých modelů.

Co výsledný graf říka o n našich modelech a je toto ověření dostatečné pro validaci model?

Navrněte další úpravy modelu za cílem co nejlépe predikvoat cenu nemovitosti.