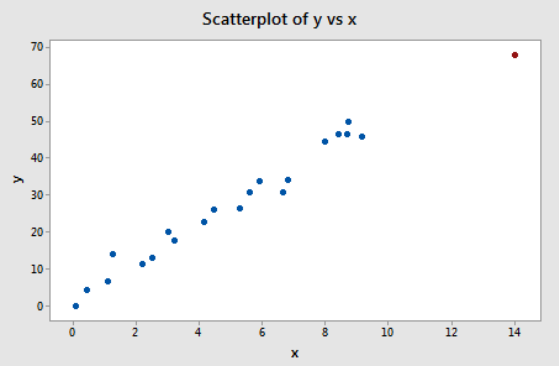
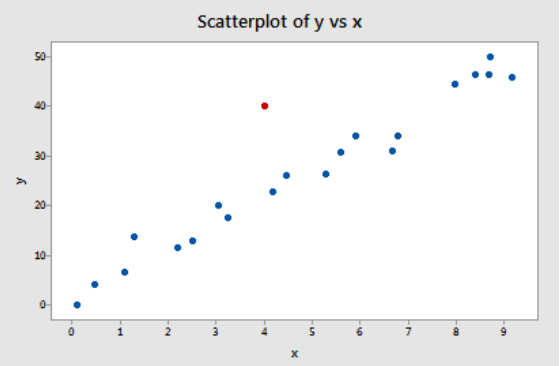
UVOD U NAUKU O PODACIMA

I kolokvijum 2020/2021

1. Učitati biblioteku **MASS** i skup podataka **Boston**. Takođe, učitati i bibiloteku **tidyverse**.
2. Grafički prikazati odnos starosti i cene kuće u odnosu na blizinu reke *Charles*.
3. Nacrtati grafik koji prikazuje raspodelu starosti kuća. Prema slobodnoj proceni dodati **novu** promenljivu frejmu koja ima nivoe starosti, npr. „nova“, „srednja i „stara“.
4. Napraviti nekoliko dijagrama raspršenosti (*scatter*). Da li se sa grafika mogu uočiti neke promenljive koje su u vezi sa cenom kuće?
5. Treba proceniti cenu kuće u zavisnosti od prediktora **X** za koji smatrate da ima najveći potencijal. Objasniti zašto je izabran baš taj neki prediktor.
6. Podeliti skup podataka na trening i test skupove u razmeri 85%:15%. **Model trenirati na trening skupu.** Napraviti model proste linearne regresije.
7. Intrepretirati dobijeni model – *RSS*, *R2*, *koeficijente regresije*, *F-statistika*.
8. Šta se može zaključiti sa grafika koji se dobijaju kao rezultat vašeg modela?
9. Da li ima potrebe raditi sa izabranim prediktorom polinomijalnu regresiju? Ako ima objasniti zašto i uraditi dodatni model. Ako nema potrebe, tada za dobijeni model iz tačke 6 uraditi predikciju nad testnim podacima.
10. Dodajte još jedan prediktor u vaš model. Naravno, opet onaj za koji smatrate da najviše obećava. 😊 Da li novi model bolje fituje podatke?
11. Da li će dodavanje starosti kuće i njene udaljenost od reke doprineti modelu?
12. Šta predstavljaju crvene tačke na slikama koje su date u prilogu?





Објаснити појам *регуларизације* (методе сакупљања, shrinkage methods) и укратко објаснити механизам рада. Шта се постиже њеним коришћењем?

Који су *основни изазови* о којима морамо водити рачуна када *је модел линеарне регресије* развијен?

Објаснити шта је *bootstrap* и чему служи.