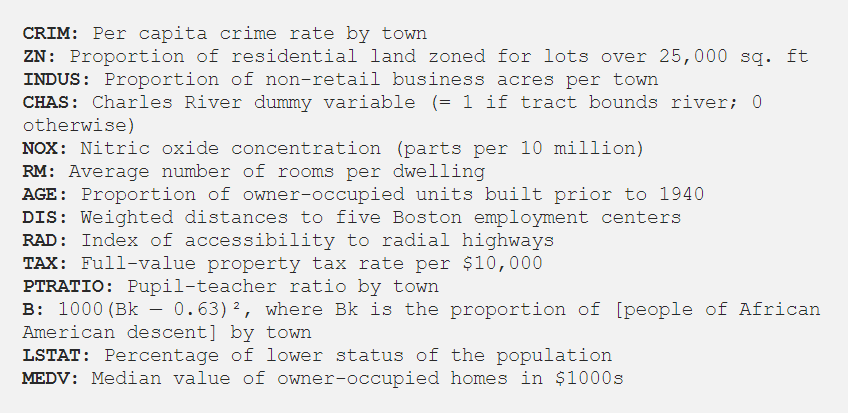
Zadatak 5 iz LV4

Varijable:



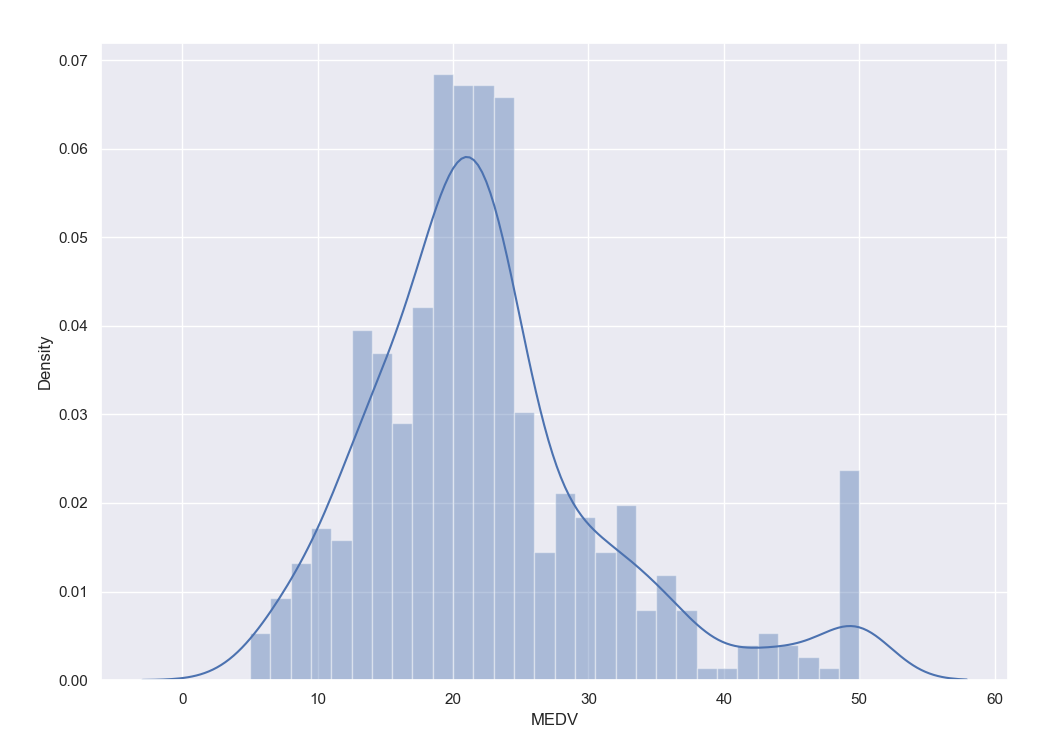
Cijene kuće označene varijablom MEDV naša su ciljna varijabla, a na temelju ostalih varijabli ćemo predvidjeti vrijednost kuće.

Nakon učitavanja podataka dobra je praksa vidjeti nedostaju li vrijednosti u podacima. Brojimo nedostajuće vrijednosti za svaku značajku koristeći isnull().

# Eksplorativna analiza podataka

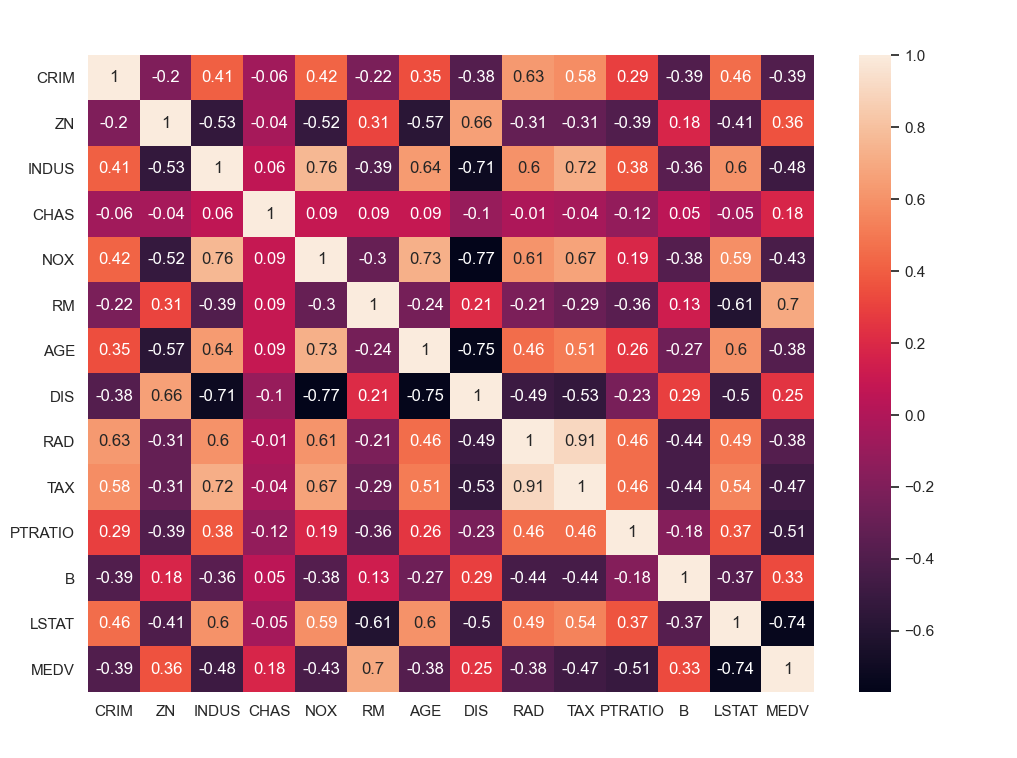
U ovom ćemo koraku koristiti neke vizualizacije kako bismo razumjeli odnos ciljne varijable s drugim značajkama.

Nacrtajmo najprije raspodjelu ciljne varijable MEDV. Koristit ćemo distplot funkciju iz seaborn knjižnice.



Vidimo da se vrijednosti MEDV distribuiraju normalno s nekoliko odstupanja.

Zatim kreiramo korelacijsku matricu koja mjeri linearne odnose između varijabli. Matrica korelacije može se oblikovati upotrebom corr funkcije iz pandas knjižnice podataka. Koristit ćemo heatmap za crtanje matrice korelacije.

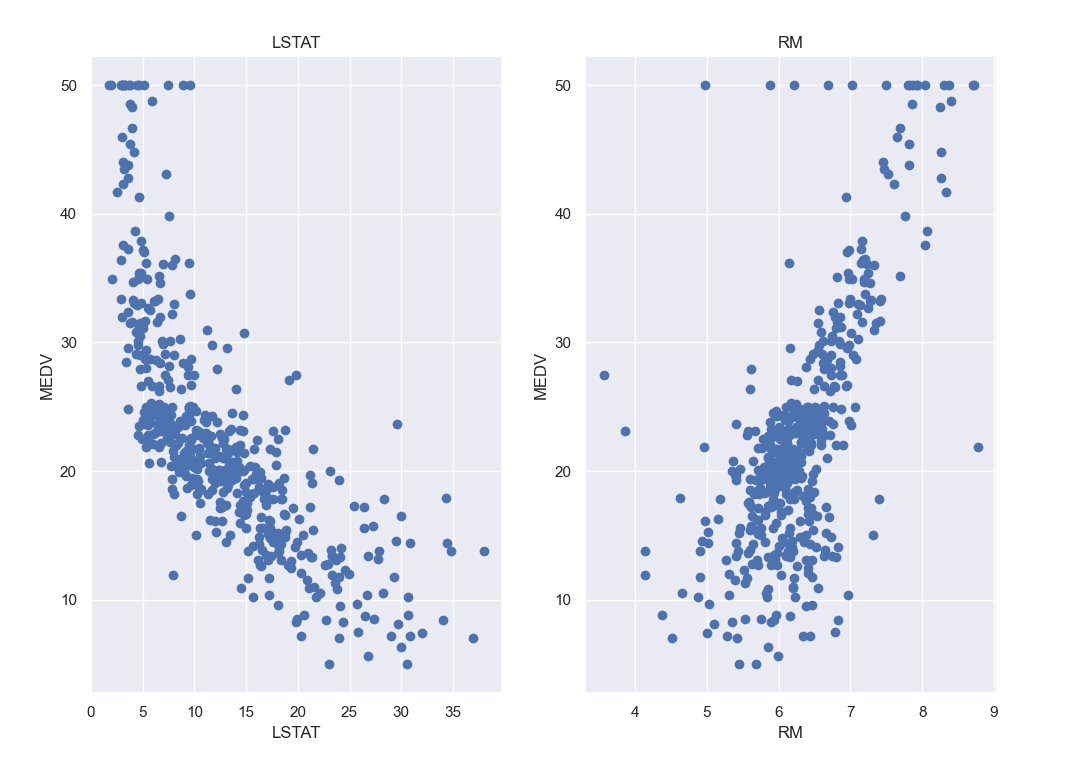


Koeficijent korelacije kreće se od -1 do 1. Ako je vrijednost blizu 1, to znači da između dvije varijable postoji snažna pozitivna korelacija. Kad je blizu -1, varijable imaju jaku negativnu korelaciju.

Da bismo uklopili model linearne regresije, odabiremo one značajke koje imaju visoku korelaciju s našom ciljnom varijablom MEDV. Gledajući korelacijsku matricu možemo vidjeti da RMima snažnu pozitivnu korelaciju s MEDV(0,7), gdje LSTATima visoku negativnu korelaciju s MEDV(-0,74).

Važna točka u odabiru značajki za model linearne regresije je provjera multi-co-linearity. Značajke RAD i TAX imaju korelaciju 0,91. Ovi su parovi značajki međusobno u korelaciji. Ne bismo trebali odabrati obje ove značajke zajedno za izradu modela. Isto vrijedi i za značajke DIS i AGE koje imaju korelaciju -0.75.

Korištenjem scatter plot ćemo prikazati kako ove značajke variraju s MEDV.



Cijene rastu kako se vrijednost RM linearno povećava. Malo je odstupanja, ali čini se da su podaci ograničeni na 50.

Cijene se obično smanjuju s porastom LSTAT-a, no izgleda da ne slijedi točnu liniju.

# Priprema podataka za obuku modela

Prvo spajamo LSTAT i RM stupce koristeći np.c iz numpy paketa.

**Podjela podataka u skup za učenje i testiranje**

Model učimo s 80% uzoraka i testiramo s preostalih 20%.

Za podjelu podataka koristimo train\_test\_split funkciju iz scikit-learn paketa.

Ispisujemo veličine našeg seta za učenje i seta za provjeru kako bismo provjerili je li razdvajanje ispravno.

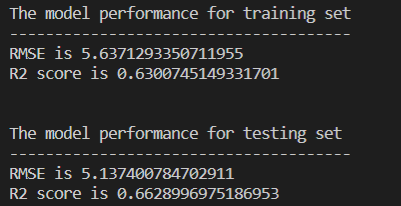


**Obuka i testiranje modela**

Koristimo LinearRegression iz scikit-learn paketa u skupu za učenje i testiranje.

# Model evaluation

Za evaluaciju modela koristit ćemo RMSE i R2-score.



Plotting y-test vs y\_predict

Idealno bi bilo da je ravna linija.

