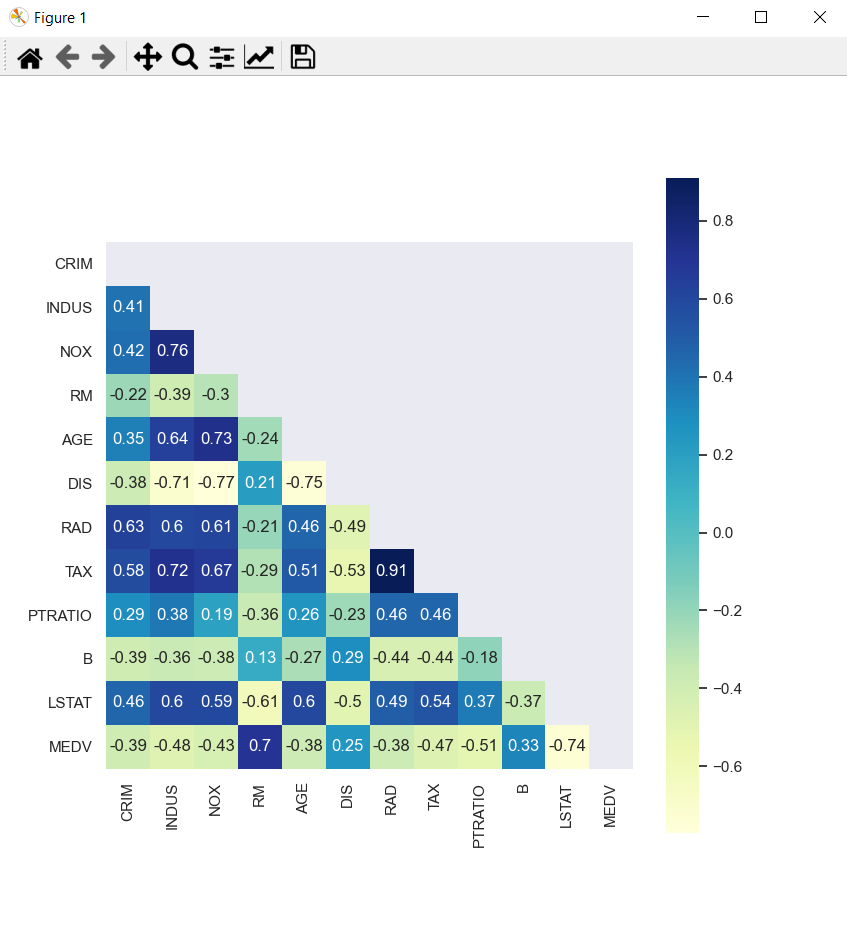
**BOSTON HOUSING PRICE PREDICTION**

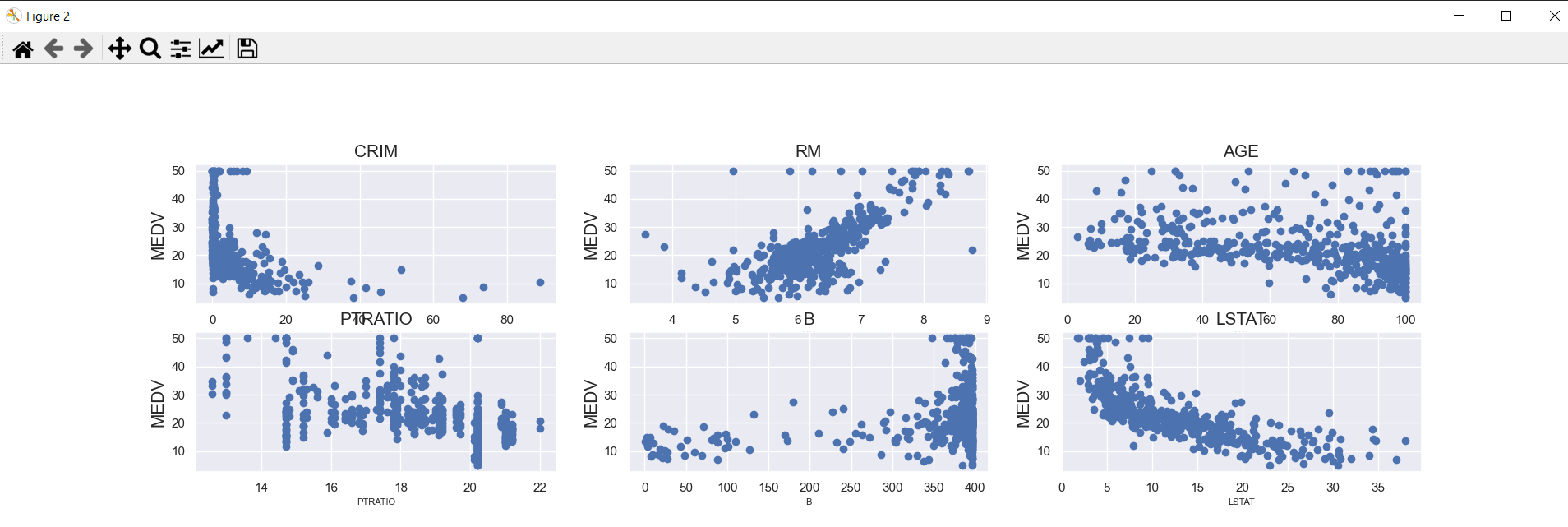
Zadatak sam započeo sa istraživanjem o samom datasheet-u i koracima koje trebam napraviti kako bih dobio sto točniji finalni graf. Sve reference ću postaviti na zadnju stranicu radi jednostavnosti. Odlučio sam se za korištenje Pandas dataframe-a pošto ga se sjećam iz prijašnje vježbe, a i na internetu se koristi u većini slučajeva izrade regresijskih modela. Kao ciljanu varijablu koristiti ću srednju vrijednost domova(MEDV- Median Value of homes u tablici). Prvi korak osim samog učitavanja tablice mi je odrediti postotak vrijednosti koje su prazne ili im je iznos 0. Ukoliko je postotak vrijednosti koje nedostaju iz stupca veći od 70% taj stupac možemo zanemariti. U našem slučaju to su stupci ZN i CHAS. Nakon toga je potrebno odrediti koji stupci su međusobno povezani kako bi izabrali one koji si najbolje povezani sa našom ciljanom varijablom. Kod za heatmapu sam našao na stackoverflow portalu i uz malo promjena dobio sam graf koji točno pokazuje ovisnosti između stupaca.

Dobiveni graf:



Iz grafa vidimo da MEDV poprilično ovisi o LSTAT stupcu.(LSTAT - % populacije koja pripada nižnoj klasi). Također uočavamo da su stupci RAD-TAX i DIS-AGE povezani. Zbog problema eng. Multicolinarity za naš projekt bi ih bilo dobro maknuti.

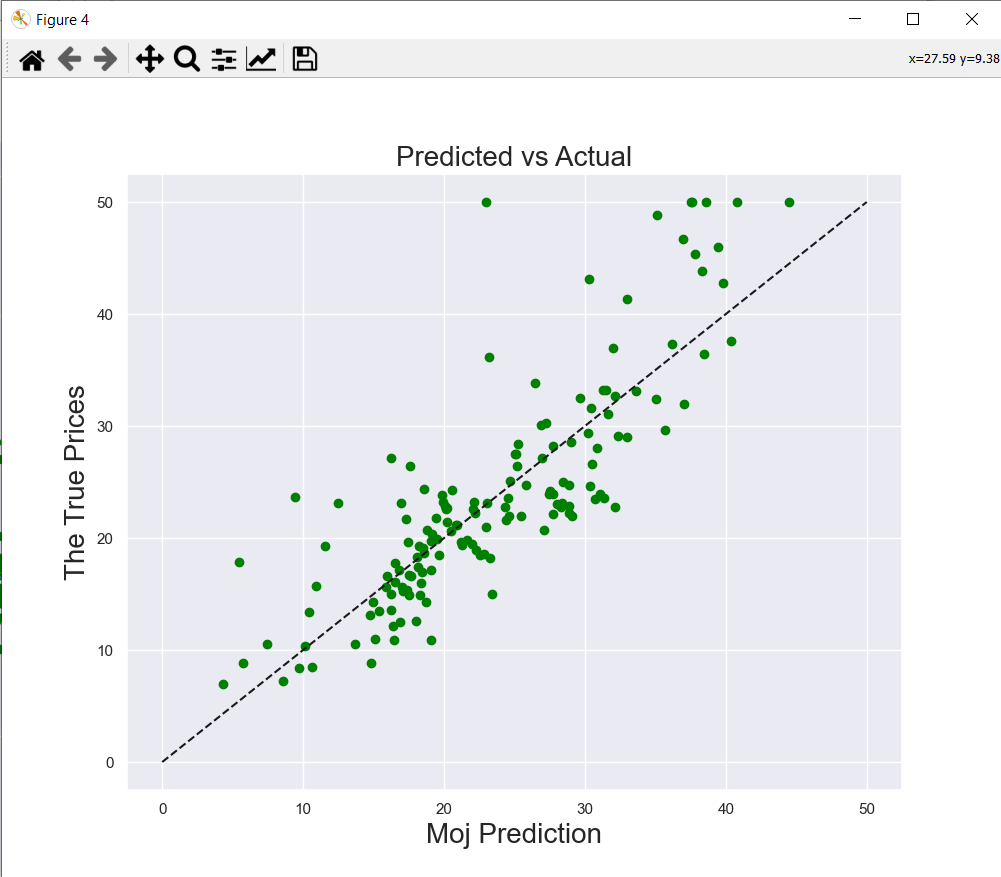
Treći korak je vizualizacija preostalih stupaca kako bi pronašli onaj koji je najbliže linearnoj ovisnosti. Prilikom izrade sam pokušao izraditi figure koji bi prikazivao sve grafove istovremeno ali tek nakon popriličnog broja pokušaja sam pronašao funkciju subplot.



Odabir grafa je bio poprilično lagan, LSTAT izgleda kao jako lijepa logaritamska funkcija koja uz pomoć komande „np.log“ postaje linearna.



Proces linearne regresije sam započeo odabirom testnih skupova, skup za učenje iznosi 70% a skup za testiranje 30%. Korištenje komanda LinearRegeression i još jednim plot-om dobivamo zadnji graf koji nam pokazuje pravu cijenu i našu predviđenu.



**ZAKLJUČAK**

Iznenađen sam samom korelacijom između stupca LSTAT i MEDV. Očekivao sam da će stupci kao PTRATIO i CRIM imati puno veći utjecaj. U svakom slučaju sam naučio dosta o samoj izradi grafova tijekom ove vježbe i načinu na koji radi sam proces pripreme podataka za linearnu regresiju.

**KOD**

<https://github.com/Mihael283/PSU_LV/blob/main/LV4/primjer_4_3.py>

**REFERENCE**

https://seaborn.pydata.org/examples/many\_pairwise\_correlations.html

https://stackoverflow.com/questions/53822194/python-generate-a-mask-for-the-lower-triangle-of-a-matrix

<https://stackoverflow.com/questions/33282368/plotting-a-2d-heatmap-with-matplotlib>

<https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html>

<https://jeffmacaluso.github.io/post/LinearRegressionAssumptions/>

https://stats.stackexchange.com/questions/1149/is-there-an-intuitive-explanation-why-multicollinearity-is-a-problem-in-linear-r/1150#1150