ITMAL Øvelser – Dataanalyse

Øvelse 1 :

I skal analysere på ”California housing prices” (https://www.kaggle.com/camnugent/california-housing-prices), som også benyttes i lærebogen.

a) Plot fordelingen af median_income. Find også spredning, middelværdi og median.
b) Er der forskel på median og middelværdi af median_income ? Hvilken af de to beskriver bedst en ”almindelig families indkomst” og hvorfor ?
c) Fit en normalfordeling til data og plot histogrammet – passer de to ?
d) Er der sammenhæng imellem median_house_value og median_income ? Lav korrelationsplot.
e) Hvad er 5% og 95% percentilerne af median_house_value ? (dvs. grænserne for 5% laveste og højeste). Plot også fordelingen af median_house_value. Kommentér på realismen af max-værdi og 95% percentil – foreslå gerne en løsning til hvad man kan gøre ved dette, hvis man skal have mere realistiske data.

Tips :

I kan finde mange eksempler på data analyse under ”Kernels” på Kaggle site, fx. - https://www.kaggle.com/rajritu2803/california-housing-price-prediction

https://www.kaggle.com/takedown/complete-tutorial-for-beginners

Seaborn tutorial – plot fordelinger og korrelationer - https://seaborn.pydata.org/tutorial/distributions.html

Øvelse 2 :

Lav data analyse på jeres egne data og projekt.