Anvendt statistik – Opgave 1

Peergrade opgave

I denne opgave vil i blive introduceret for Kickstarter data. Kickstarter er en amerikansk internetplatform, hvor virksomheder og start-ups kan søge crowd-funding til interessante projekter og ideer.

Igennem denne opgave vil i blive bedt om at anvende Python (fx i Colab eller Anaconda) til at opnå indsigt i Kickstarter, med udgangspunkt i de funktioner og formler i er blevet introduceret til gennem undervisning og øvelser.

For at indlæse data, skal i bruge en URL, som vist i tidligere arbejdseksempler fra undervisning (både Titanic og Pingvindata). Til Kickstarter opgaven får i her udleveret link der giver adgang til data.
Kickstarter URL: 'https://sds-aau.github.io/IntroStat/Data/kickstarter.xlsx' (**Hint:** bruges sammen med pandas read_excel funktionen)

1. Indlæsning af data:

A. Start med at hente de pakker i skal bruge.
   Note: Dette kan også ske løbende, men for jeres eget overbliks skyld, vil det være godt at gøre fra starten.
B. Indlæs jeres data.
   Note/hint: I skal bruge pandas funktionen .read_excel(), som vist i Notebooks fra undervisning

2. Første databehandling

A. Vis de 5 første rækker af data.
B. Giv et overblik over dataets informationer (brug indbygget funktion).
C. Drop manglende værdier på tværs af datasættet (dvs. hvis der mangler én variable for en observation, skal hele observationen ud). Hvor mange observationer er der nu tilbage?

3. Frekvensberegninger & nøgletal

A. Giv et overblik over hvor mange forskellige kategorier (main_category) der er på Kickstarter? Lav en oversigt over hvor mange projekter der er i de 10 mest populære kategorier.

Hint: Beregn først hvor mange der er i alle kategorier og så kan du bruge indexing til at vise de første 10 → `objekt[:10]` (mere om det her: https://jakevdp.github.io/PythonDataScienceHandbook/03.02-data-indexing-and-selection.html)

Kan du se hvilken kategori der er mest populær?

B. Giv et overblik over hvor mange lande der er repræsenteret på Kickstarter
   Hvor mange danske projekter er der på Kickstarter?
C. Hvor meget funding søger projekterne i gennemsnit på tværs af platformen?
   Beregn 1. og 3. kvartil for samme variable.
   Hint: brug variablen ”goal” til at finde det gennemsnitlige indsamlingsmål
D. Giv et overblik over hvor meget de forskellige kategorier i gennemsnit søger i funding og vise de top10 kategorier.
   Hint: Her kan der tilføjes ” .sort_values(ascending=False)” for at sortere fra højeste til laveste.
   Lav den samme beregning for median.
   Bonus: Hvilken kategori søger i gennemsnit/median mest funding, og hvor meget? Der er ret store forskelle mellem median og gennemsnit her. Har du nogle ideer om hvorfor?
E. Hvor mange støtter (brug variablen ”backers”) har Kickstarter projekterne? Brug nøgletal til at beskrive gennemsnit, kvartiler og min- og maxværdier på tværs af samtlige projekter. Hvad kan du sige om den her fordeling? Hvad siger det (muligvis) om de forskellige projekter?

4. Visualisering (ikke nemt ☺)

A. Histogram over ‘**usd_pledged_real’**

B. Boxplot, der viser fordelinger for ‘**usd_pledged_real’** vs success (**’state’**)

C. En udgave af datasættet uden outliers (værdier for **usd_pledged_real** over 90 percentil) og plot de to grafer igen.

D. Fortolk graferne.