

Dadansoddi Data gyda Python ac R

Dr. Geraint Palmer

1 Cyflwyniad

Mae defnyddio meddalwedd ystadegol arbenigol er mwyn gwneud dadansoddi data yn helpu cywirdeb ac effeithlonrwydd eich gwaith. Yn y cwrs hwn byddwn yn edrych ar naill ai R neu Python (eich dewis chi), dwy iaith rhaglennu boblogaidd iawn ar gyfer dadansoddi data a thasgau cysylltiedig. Mae defnyddio cod, ac yn enwedig cod ffynhonnell agored, ar gyfer hwn hefyd yn helpu gyda ailgynhyrchadwyedd - mae mwy o wybodaeth ar y fath o gysyniadau hyn: <https://sgiliauymchwilcyfrifiadurol.github.io/>.

2 Cyfarwyddiadau

Dewiswch i ddilyn y cwrs hwn yn defnyddio naill ai Python neu R. Dewiswch yr iaith sydd fwyaf priodol i'ch sefyllfa chi - os ydych yn gweld eich bod yn fwy tebygol o ddefnyddio un dros y llall yn eich gyrfa ymchwil dewiswch yr un hynny. Os nad oes gennych ddewis amlwg, darllenwch yr adran 'Python ac R' isod, cymerwch gipolwg ar y fideos i weld sut maent yn edrych, ac ymchwiliwch ar-lein er mwyn gwneud penderfyniad.

Mae pump fideo'r un ar gyfer pob iaith. Gwylwch y fideos ar gyfer yr iaith gwnaethoch chi ddewis. Pynciau'r fideos yw:

1. Lawrlwytho & chystrawen syml
2. Darllen i mewn data & ystadegau disgrifiadol
3. Plotio
4. Cyfuno & Ail-siapiro data
5. Profi rhagdybiaethau

Mae'r data rydym yn defnyddio yn y fideos ar gael i chi lawrlwytho, felly rydw i'n argymhell i chi ceisio ail-greu'r dadansoddiad yn y fideos.

Yna, cymhwyswch yr hyn rydych wedi dysgu trwy weithio trwy'r tasgau isod.

Bydd cyfle i drafod eich datrysiadau ac unrhyw un o'r cysyniadau hyn yn y sesiwn fyw.

3 Tasgau

Gan ddefnyddio'r data `data_tai.csv`:

1. Beth yw pris cymedrig y tai yn Aberwylan ac yn Llanceiliog?
2. Ar gyfartaledd, plant o Aberwylan neu Llanceiliog sydd angen cerdded y pellter mwyaf i'r ysgol?
3. Crëwch dabl o faint o dai o bob maint sydd ym mhob tref.
4. Crëwch blotiau gwasgariad o'r pellteroedd o'r ysgol a'r eglwys yn erbyn pris y tŷ.

5. Crëwch histogram o brisiau'r tai.
6. Crëwch blotiau bocs ochr yn ochr o brisiau tai yn ôl maint.
7. Gan ystyried y data fel sampl o holl dai'r ddwy dref, defnyddiwch brawf ystadegol priodol i weld os oes wahaniaeth rhwng prisiau tai mawr canolog a phrisiau tai bach.
8. Gan ystyried y data fel sampl o holl dai'r ddwy dref, defnyddiwch brawf ystadegol priodol i weld os oes wahaniaeth rhwng prisiau tai Aberwylan a phrisiau tai Llanceiliog.

4 Python ac R

Mae Python ac R yn ddwy iaith rhaglennu ffynhonnell agored. Mae hwn yn golygu cwpl o bethau, mae'r cod ffynhonnell i gyd ar gael i ni edrych arno, a newid os oes angen. I ni fel defnyddwyr y prif beth mae hwn yn golygu yw bod ganddynt drwydded agored, a gallwn ni lawrlwytho a defnyddio nhw am ddim. Mae hefyd ganddynt gymuned o ddatblygwyr sy'n adeiladu ac yn cynnal pecynnau neu lyfrgelloedd defnyddiol, am ddim, ar eu cyfer. Fel arfer wrth ysgrifennu cod byddwn yn defnyddio nifer o'r pecynnau/llyfrgelloedd hyn yn rheolaidd.

Fan hyn byddaf yn disgrifio'n fras y ddwy iaith. Er mwyn rhagatal unrhyw fias yn yr ysgrifennu isod, rydw i fy hun yn defnyddio Python pob dydd. Ond nid yw hwn yn golygu fod un iaith yn well na'r llall.

Fel arfer gwelwn R fel iaith sy'n arbenigo mewn dadansoddi data ac ystadegaeth. Mae ei gystrawen wedi ysgrifennu i weithio'n naturiol gyda data a matricesau. Mae rhan fawr o'i gymuned o ddefnyddwyr yn cynnwys ystadegwyr a gwyddonwyr eraill, ac y rhain sy'n datblygu a chynnal y llwyth o becynnau defnyddiol ar gyfer R. O ganlyniad, mae nifer fawr o becynnau arbenigol ar gyfer ystadegaeth, wedi'i adeiladu gan ystadegwyr neu'r gwyddonwyr sy'n ei defnyddio. Fel arfer caiff papurau ystadegol academiaidd newydd ei throsi'n becyn R yn fuan, sy'n golygu fod pecynnau arbenigol iawn, er efallai bod nifer o becynnau neu nifer o ffyrdd gwahanol o wneud yr un dasg.

Mae Python yn iaith rhaglennu fwy cyffredinol, sy'n cael ei ddefnyddio ar gyfer ystod eang o gymwysiadau megis adeiladu gwefannau, dadansoddi data, dysgu peirianyddol, a chreu gemau cyfrifiadurol. Cafodd ei dylunio i fod yn narllenadwy. Mae ei gymuned yn cynnwys amrywiad mawr, gan gynnwys gwyddonwyr, dadansoddwyr dysgu peirianyddol, a datblygwyr we. Felly mae ystod eang o wahanol fathau o lyfrgelloedd poblogaidd iawn ganddo er mwyn gwneud nifer fawr o dasgau gwahanol, sydd fel arfer yn hawdd i'w gyfuno. Mae hwn yn golygu fod llyfrgelloedd Python yn gallu dod yn boblogaidd iawn, ac mae tuedd o ychwanegu at a gwella'r llyfrgelloedd poblogaidd hyn, felly mae'r rhain yn dod yn ffyrdd 'safonol' o wneud rhyw dasg yn Python.

Mewn gwirionedd, ar gyfer y fath o dasgau byddwn yn edrych arnynt yn y cwrs hwn, does fawr o wahaniaeth rhwng y ddau. Y prif beth i ystyried yw eu hecosystemau, hynny yw'r amrywiaeth a lefel arbenigedd y pecynnau a'r llyfrgelloedd sydd ar gael amdanynt. Mae gan y ddau cymunedau gwych, gyda nifer fawr o gynadleddau, a grwpiau defnyddwyr lleol, a defnyddwyr sy'n fywiog iawn ar fforymau ar-lein.