Dadansoddi Data gyda Python

Cynnwys

1 Cystrawen cyffredinol

Mae *newidynnau* yn pwyntio at wrthrychau yn Python. Gallen nhw fod yn rhifau, llinynnau, math Boole, rhestrau, ac yn y blaen. Gallwn gwneud rhifyddeg ar rhain:

```
>>> fy_oedran = 32
>>> fy_oedran_blwyddyn_nesaf = fy_oedran + 1
>>> fy_oedran_blwyddyn_nesaf
33

>>> ydw_in_blentyn = fy_oedran < 18
>>> ydw_in_blentyn
False

>>> fy_enw = 'Geraint'
>>> fy_cyfenw = 'Palmer'
>>> fy_enw_llawn = fy_enw + fy_cyfenw
>>> fy_enw_llawn
'GeraintPalmer'

>>> fy_hoff_bethau = [3.14159, 'pinafal', 22, True, 'mathemateg']
>>> fy_hoff_bethau[1]
'pinafal'
```

Mae *ffwythiannau* yn gallu allbynnu gwethoedd o werthoedd eraill, fel yn mathemateg:

```
>>> def sgwario(x):
...     return x ** 2
>>> sgwario(5)
25
```

Mae *llyfrgellau* yn galluogi ni i ddefnyddio newidynnau a ffwythiannau y mae pobl eraill wedi ysgrifennu:

```
>>> import math
>>> math.pi
3.141592653589793
>>> math.sin(2.2)
0.8084964038195901
```

2 Fframiau data

Llyfrgell pwysig ar gyfer dadansoddi data mewn Python yw `pandas`, sy'n galluogi ni i ddelio gyda fframiau data. Gallwn darllen i mewn data:

```
>>> import pandas as pd
>>> data = pd.read_csv('data.csv')
>>> data
```

	Enw	Rhyw	Oedran	Hyd Aros yn yr Ysbyty	Pwysau Dechreuol	Math
0	Alun Adams	G	72	2.5	82.25	Anaf ffisegol
1	Bleddyn Bowen	G	72	2.0	81.09	Anaf ffisegol
2	Christopher Clwyd	G	87	3.0	81.78	Anaf ffisegol
3	Daniel Derwen	G	89	6.5	81.29	Anaf ffisegol
4	Eifion Evans	G	96	15.5	75.17	Problem meddygol
5	Frank Fillmore	G	58	8.0	78.92	Anaf ffisegol
6	Abigail Ashcroft	B	97	1.5	74.26	Anaf ffisegol
7	Beryl Bones	B	71	6.5	68.67	Problem meddygol

Gallwn defnyddio cromfachau sgwâr er mwyn cael colofn penodol:

```
>>> data['Oedran']
0     72
1     72
2     87
3     89
4     96
...
30    75
31    92
32    61
33    69
34    65
Name: Oedran, dtype: int64
```

Mae yna dulliau gallwn ddefnyddio ar y colofnau hyn sy'n rhoi ystadegu disgrifiadol o'r colofn:

```
>>> data['Oedran'].mean()
76.31428571428572

>>> data['Oedran'].median()
75.0

>>> data['Oedran'].max()
97

>>> data['Oedran'].min()
48

>>> data['Oedran'].var()
148.3983193277311
```

Gallwn hidlo fframdata trwy ofyn ond am y rhesi sydd yn bodloni rhyw amod. Er enghraifft, i cael ond y cleifion sydd ag oedran yn fwy nag 89:

```
>>> data[data['Oedran'] > 89]
```

	Enw	Rhyw	Oedran	Hyd Aros yn yr Ysbyty	Pwysau Dechreuol	Math
3	Daniel Derwen	G	89	6.5	81.29	Anaf ffisegol
4	Eifion Evans	G	96	15.5	75.17	Problem meddygol
6	Abigail Ashcroft	B	97	1.5	74.26	Anaf ffisegol
8	Carys Carnegie	B	91	2.5	72.12	Problem meddygol
18	Irene Innes	B	90	4.0	67.01	Problem meddygol
20	Larry Lawrence	G	92	3.0	71.71	Anaf ffisegol
31	Renne Rutherford	G	92	1.5	81.45	Anaf ffisegol

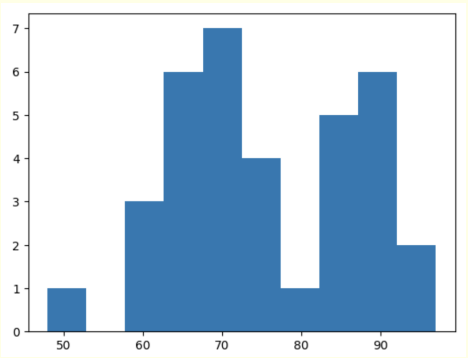
Ac felly gallwn cael ystadegau disgrifiadol o'r fframdata newydd hwn. Pwysau dechrauol cymedrig y cleifion sydd yn hynach nag 89 mlwydd oed yw:

```
>>> data_dros_89 = data[data['Oedran'] >= 89]
>>> data_dros_89['Pwysau Dechreuol'].mean()
74.71571428571428
```

3 Plotio

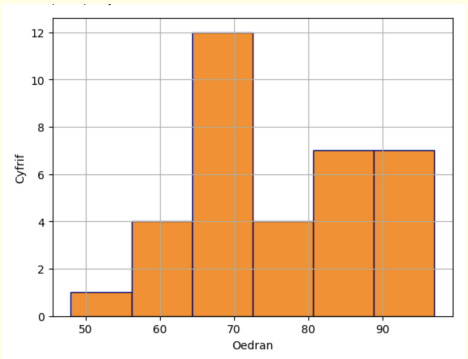
Gallwn creu delweddau er mwyn deall neu cyfathrebu'r data yn haws trwy'r llyfrgell `matplotlib`. Mae nifer fawr o wahanol fathau o plotiau gallwn creu, a gallwn eu creu trwy rhoi rhestr(au) o rhifau, neu colofn(au) o fram data i mewn i ffwythiannau o matplotlib. Er enghraifft, histogram o oedranm cleifion:

```
>>> import matplotlib.pyplot as plt
>>> plt.hist(data['Oedran']);
```



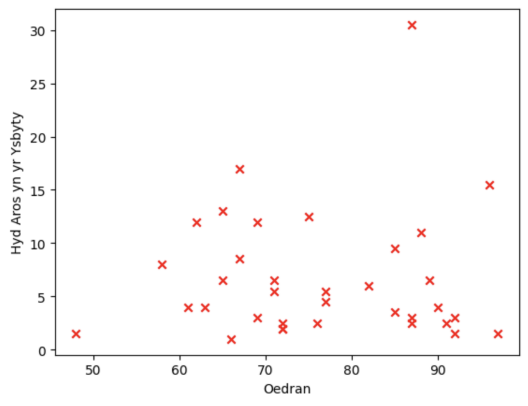
Mae'r llyfrgell hefyd yn galluogi addasiadau esthetig i'r diagram, megis lliwiau a labelau. Er enghraifft:

```
>>> plt.hist(
...     data['Oedran'],
...     facecolor='darkorange',
...     edgecolor='darkblue',
...     bins=6
... )
>>> plt.grid()
>>> plt.xlabel("Oedran")
>>> plt.ylabel("Cyfrif")
```



Mae rhai plotiau angen mwy nag un colofn o ddata, er enghraifft plot gwasgariad:

```
>>> plt.scatter(
...     data['Oedran'],
...     data['Hyd Aros yn yr Ysbyty'],
...     marker='x',
...     c='red'
... )
>>> plt.xlabel('Oedran')
>>> plt.ylabel('Hyd Aros yn yr Ysbyty')
```



4 Cyfuno & ail-siapio

Mae ail-siapio data yn ddefnyddiol, y ffordd symlaf yw trwy'r dull `value_counts`, sy'n cyfri faint o rhesi sydd gan pob werth o rhyw colofn:

```
>>> data['Math'].value_counts()
Math
Anaf ffisegol       24
Problem meddygol    11
Name: count, dtype: int64
```

Ffordd arall o ail-siapio data yw trwy'r dull `groupby`, sy'n grwpio data yn ôl gwerth rhyw colofn. Mae hwn yn rhoi gwrthrych sy'n ymddwyn fel fram data, ond ni allwn ei weld nes i ni defnyddio dulliau arno. Creu'r gwrthrych:

```
>>> grwpio_gan_rhyw = data.groupby('Rhyw')
>>> grwpio_gan_rhyw
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x11f430700>
```

Nawr wrth ddefnyddio dulliau Pandas arno, mae'r dulliau hyn yna yn cael eu cymhwyso ar pob grŵp ar wahân. Er enghraifft:

```
>>> grwpio_gan_rhyw['Oedran'].mean()
Rhyw
B    72.666667
G    79.050000
Name: Oedran, dtype: float64

>>> grwpio_gan_rhyw['Math'].value_counts()
Rhyw  Math
B     Problem meddygol     8
      Anaf ffisegol        7
G     Anaf ffisegol       17
      Problem meddygol     3
Name: count, dtype: int64
```

Gallwn cyfuno data gyda `pd.merge` a `pd.concat`. Ystyriwch pan fod gennym colofn ychwanegol gyda'r un rhesi a'n data:

```
>>> colofn_ychwanegol = pd.read_csv("colofn_ychwanegol.csv")
>>> colofn_ychwanegol
```

	Enw	Pwysau wrth Gadael
0	Alun Adams	79.21
1	Bleddyn Bowen	79.06
2	Christopher Clwyd	78.91
3	Daniel Derwen	80.12
4	Eifion Evans	72.81
5	Frank Fillmore	76.81
6	Abigail Ashcroft	74.43

Cyfunwn rhain gyda `pd.merge`:

```
>>> data_gyda_colofn_ychwanegol = pd.merge(data, colofn_ychwanegol)
>>> data_gyda_colofn_ychwanegol
```

	Enw	Rhyw	Oedran	Hyd Aros yn yr Ysbyty	Pwysau Dechreuol	Math	Pwysau wrth Gadael
0	Alun Adams	G	72	2.5	82.25	Anaf ffisegol	79.21
1	Bleddyn Bowen	G	72	2.0	81.09	Anaf ffisegol	79.06
2	Christopher Clwyd	G	87	3.0	81.78	Anaf ffisegol	78.91
3	Daniel Derwen	G	89	6.5	81.29	Anaf ffisegol	80.12
4	Eifion Evans	G	96	15.5	75.17	Problem meddygol	72.81
5	Frank Fillmore	G	58	8.0	78.92	Anaf ffisegol	76.81
6	Abigail Ashcroft	B	97	1.5	74.26	Anaf ffisegol	74.43

Ac pan fod gennym rhesi ychwanegol gyda'r un colofnau a'n data:

```
>>> rhesi_ychwanegol = pd.read_csv("rhesi_ychwanegol.csv")
>>> rhesi_ychwanegol
```

	Enw	Rhyw	Oedran	Hyd Aros yn yr Ysbyty	Pwysau Dechreuol	Pwysau wrth Gadael	Math
0	Pamela Potts	B	63	8.0	70.32	70.06	Problem meddygol
1	Quinlan Queltch	B	80	8.5	75.34	77.12	Anaf ffisegol
2	Ulysses Underhill	G	65	12.5	86.47	82.12	Anaf ffisegol
3	Victor Vallance	G	81	6.0	76.66	72.42	Problem meddygol
4	Rosa Roberts	B	78	3.0	79.82	74.13	Problem meddygol

Cyfunwn rhain gyda `pd.concat`:

```
>>> data_gyda_rhesi_ychwanegol = pd.concat(
...     [rhesi_ychwanegol, data_gyda_colofn_ychwanegol]
... )
>>> data_gyda_rhesi_ychwanegol
```

	Enw	Rhyw	Oedran	Hyd Aros yn yr Ysbyty	Pwysau Dechreuol	Pwysau wrth Gadael	Math
0	Pamela Potts	B	63	8.0	70.32	70.06	Problem meddygol
1	Quinlan Queltch	B	80	8.5	75.34	77.12	Anaf ffisegol
2	Ulysses Underhill	G	65	12.5	86.47	82.12	Anaf ffisegol
3	Victor Vallance	G	81	6.0	76.66	72.42	Problem meddygol
4	Rosa Roberts	B	78	3.0	79.82	74.13	Problem meddygol
0	Alun Adams	G	72	2.5	82.25	79.21	Anaf ffisegol
1	Bleddyn Bowen	G	72	2.0	81.09	79.06	Anaf ffisegol
2	Christopher Clwyd	G	87	3.0	81.78	78.91	Anaf ffisegol
3	Daniel Derwen	G	89	6.5	81.29	80.12	Anaf ffisegol
4	Eifion Evans	G	96	15.5	75.17	72.81	Problem meddygol
5	Frank Fillmore	G	58	8.0	78.92	76.81	Anaf ffisegol
6	Abigail Ashcroft	B	97	1.5	74.26	74.43	Anaf ffisegol

5 Profi rhagdybiaethau

Mae gan y llyfrgell `scipy.stats` nifer ffwythiannau er mwyn rhedeg profion rhagdybiaeth. Yn union fel y llyfrgell plotio, mae'r ffwythiannau hyn yn cymryd rhestr(au) o rhifau, neu colofn(au) o fram data. Rhai ffwythiannau:

- `scipy.stats.ttest_1samp`: Prawf-t un sampl, yn cymharu cymedr i rhif.
- `scipy.stats.ttest_ind`: Prawf-t dau sampl, yn cymharu dau cymedr.
- `scipy.stats.wilcoxon`: Prawf Wilcoxon, yn cymharu canolrif i sero.
- `scipy.stats.mannwhitneyu`: Prawf Mann-Whitney-U, yn cymharu dau canolrif.
- `scipy.stats.f_oneway`: Prawf ANOVA un fford, yn cymharu mwy na dau cymedr.
- `scipy.stats.kruskal`: Prawf Kruskal Wallis, yn cymharu mwy na dau canolrif.

Er enghraifft:

Os ydym yn trin y data fel sampl, gallwn ofyn a yw'r oedran cymedrig yn halaf i 73 neu peidio. Fel prawf rhagdybiaeth:

- H_0: $\mu = 73$,
- H_1: $\mu \neq 73$.

Yn perfformio'r prawf ar y lefel 95%:

```
>>> import scipy.stats
>>> scipy.stats.ttest_1samp(data['Oedran'], 73)
TtestResult(statistic=1.6095685434860927, pvalue=0.116739124899234, df=34)
```

A chawn gwerth-p o 0.1167, sydd yn fwy nag $\alpha = 0.05$, ac felly does dim digon o dystiolaeth gyda ni i wrthod H_0.

Enghraifft arall:

Os ydym yn trin y data fel sampl, gallwn ofyn a yw'r hyd aros yn yr ysbty cymedrig yn wahanol ar gyfer dynion a menywod. Fel prawf rhagdybiaeth:

- H_0: $\mu_{\text{dynion}} = \mu_{\text{menywod}}$,
- H_1: $\mu_{\text{dynion}} \neq \mu_{\text{menywod}}$,

Yn perfformio'r prawf ar y lefel 95%:

```
>>> import scipy.stats
>>> scipy.stats.ttest_ind(
...     data[data['Rhyw'] == 'B']['Pwysau Dechreuol'],
...     data[data['Rhyw'] == 'G']['Pwysau Dechreuol']
... )
TtestResult(statistic=-9.05079585874, pvalue=1.8500072104976e-10, df=33.0)
```

A chawn gwerth-p o bach iawn, sydd yn llai nag $\alpha = 0.05$, ac felly gallwn wrthod H_0 a derbyn H_1.

6 Rhagor o wybodaeth

Am ragor o ddeunyddiau cyfrwng Cymraeg ar ddefnyddio Python:

- Cyfres o fideo tiwtorialau cyffredinol ar Python: `https://www.geraintianpalmer.org.uk/teaching/tiwtorialau-python/`
- Cyfres o fideo tiwtorialau ar dadansoddi data gydag R a Python: `https://www.porth.ac.uk/cy/collection/dadansoddi-data-gydag-r-a-python`
- Gwybodaeth ar sgiliau ymchwil ailgynhyrchiadwy: `https://sgiliauymchwilcyfrifiadurol.github.io/`

Ymarferion

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.