700+ Free Datasets in Python

that you can use today...

Deepnote

ON THE MARK DATA

Deepnote

🤯 700+ Free Datasets in Python

Use the package `pydataset` to get instant access!

```
import sys
!{sys.executable} -m pip install pydataset
```

A quick analyses to show you the power of `pydataset`:

```
import pydataset
import pandas as pd
```

All the datasets avaialble via `pydata`

```
pydataset.data()
```

Visualize

	dataset_id object	title object	
	npk 0.3% swiss 0.3% 725 others 99.5%	Seven data sets showing a bifactor solution. 1.1% Individual Preferences Over Immigration Pol... 0.8% 700 others 98.2%	
0	AirPassengers	Monthly Airline Passenger Numbers 1949-1960	
1	BJsales	Sales Data with Leading Indicator	
2	BOD	Biochemical Oxygen Demand	
3	Formaldehyde	Determination of Formaldehyde	
4	HairEyeColor	Hair and Eye Color of Statistics Students	
5	InsectSprays	Effectiveness of Insect Sprays	
6	JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share	
7	LakeHuron	Level of Lake Huron 1875-1972	
8	LifeCycleSavings	Intercountry Life-Cycle Savings Data	
9	Nile	Flow of the River Nile	

757 rows, showing 10 per page Page 1 of 76

Deepnote

All the datasets with Month or Quarter in Title

```
possible_data = pydataset.data()
possible_data[possible_data['title'].str.contains('Quarter|Month')]
```

	dataset_id object	title object
	AirPassengers 10% JohnsonJohnson 10% 8 others 80%	Monthly Airline Passenger Numbers 1949-1960 10% Quarterly Earnings per Johnson & Johnson Share 10% 8 others 80%
0	AirPassengers	Monthly Airline Passenger Numbers 1949-1960
6	JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share
17	UKgas	UK Quarterly Gas Consumption
30	austres	Quarterly Time Series of the Number of Australian Residents
49	nottem	Average Monthly Temperatures at Nottingham, 1920-1939
53	presidents	Quarterly Approval Ratings of US Presidents
61	sunspot.month	Monthly Sunspot Data, from 1749 to "Present"
63	sunspots	Monthly Sunspot Numbers, 1749-1983
71	acme	Monthly Excess Returns
468	deaths	Monthly Deaths from Lung Diseases in the UK

10 rows, showing 100 per page

Page 1 of 1

Deepnote

All the datasets with Month or Quarter in Title

```
possible_data = pydataset.data()
possible_data[possible_data['title'].str.contains('Quarter|Month')]
```

	dataset_id object	title object
	AirPassengers 10% JohnsonJohnson 10% 8 others 80%	Monthly Airline Passenger Numbers 1949-1960 10% Quarterly Earnings per Johnson & Johnson Share 10% 8 others 80%
0	AirPassengers	Monthly Airline Passenger Numbers 1949-1960
6	JohnsonJohnson	Quarterly Earnings per Johnson & Johnson Share
17	UKgas	UK Quarterly Gas Consumption
30	austres	Quarterly Time Series of the Number of Australian Residents
49	nottem	Average Monthly Temperatures at Nottingham, 1920-1939
53	presidents	Quarterly Approval Ratings of US Presidents
61	sunspot.month	Monthly Sunspot Data, from 1749 to "Present"
63	sunspots	Monthly Sunspot Numbers, 1749-1983
71	acme	Monthly Excess Returns
468	deaths	Monthly Deaths from Lung Diseases in the UK

10 rows, showing 100 per page

Page 1 of 1

Deepnote

We found our time series data!

Question:

Are lung deaths in the UK correlated with UK gas utilization?

```
uk_lung_deaths_monthly = pydataset.data('deaths')
uk_lung_deaths_monthly.head(3)
```

	time float64	deaths int64
1	1974.0	3035
2	1974.0833333333	2552
3	1974.1666666667	2704

3 rows, showing 100 per page Page 1 of 1

```
uk_gas_quarterly = pydataset.data('UKgas')
uk_gas_quarterly.head(3)
```

	time float64	UKgas float64
1	1960.0	160.1
2	1960.25	129.7
3	1960.5	84.8

3 rows, showing 100 per page Page 1 of 1

Deepnote

Prepare the data...

```
# turn month to quarters
uk_lung_deaths_monthly['is_quarter'] = uk_lung_deaths_monthly['time'] % 0.25
uk_lung_deaths_quarterly = uk_lung_deaths_monthly[
    uk_lung_deaths_monthly['is_quarter']==0
    ]
uk_lung_deaths_quarterly = uk_lung_deaths_quarterly[['time', 'deaths']]

uk_lung_deaths_quarterly.head()
```

	time float64	deaths int64
1	1974.0	3035
4	1974.25	2554
7	1974.5	1721
10	1974.75	2074
13	1975.0	2933

5 rows, showing 10 per page

Page 1 of 1

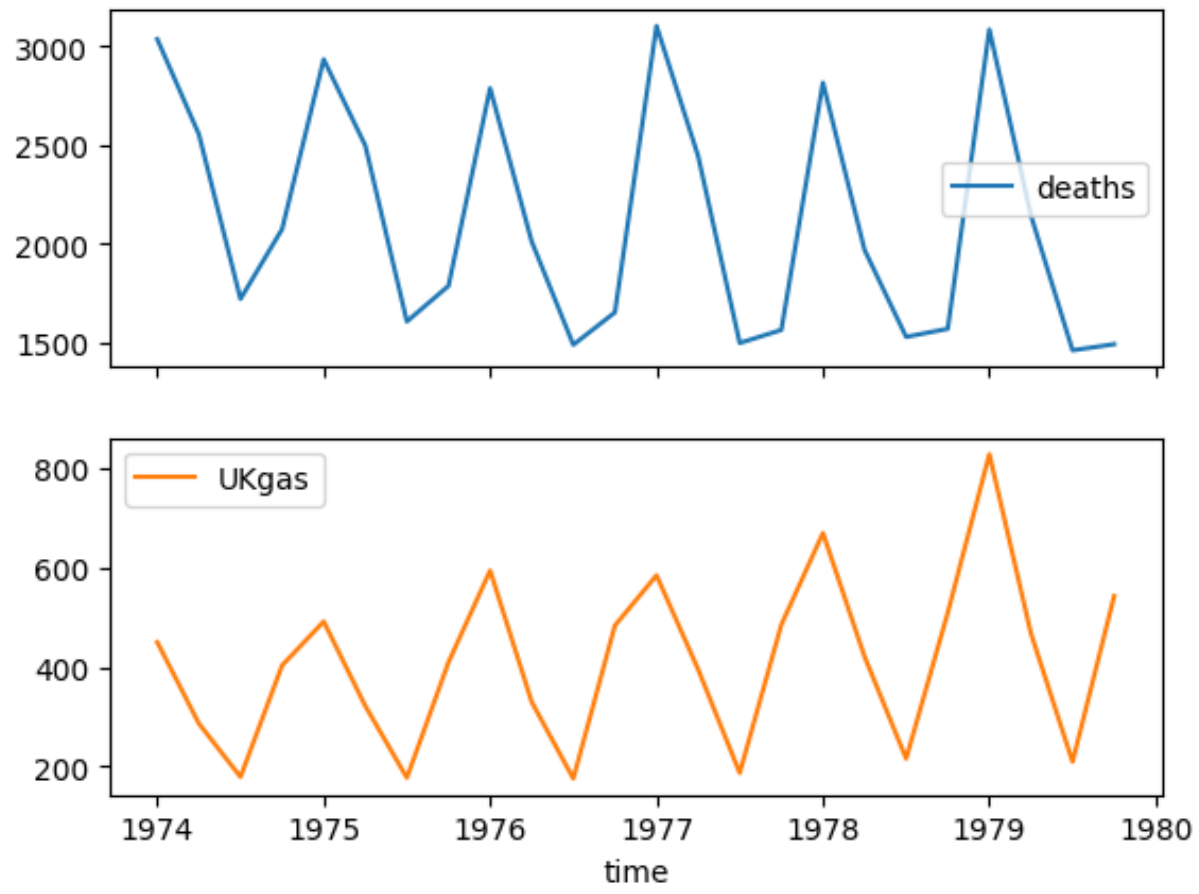
```
# get unique times
uk_lung_deaths_time = uk_lung_deaths_quarterly['time'].tolist()
uk_gas_time = uk_gas_quarterly['time'].tolist()
combined_time = uk_lung_deaths_time + uk_gas_time
unique_time = set(combined_time)
unique_time_df = pd.DataFrame(unique_time, columns=['time'])

# get max and min times
min_time = max(
    min(uk_lung_deaths_time),
    min(uk_gas_time)
    )
max_time = min(
    max(uk_lung_deaths_time),
    max(uk_gas_time)
    )

# left join data
join_df = unique_time_df \
    .merge(uk_lung_deaths_quarterly, on='time', how='left') \
    .merge(uk_gas_quarterly, on='time', how='left')

# filter and organize data
join_filter_df = join_df[
    (join_df['time'] >= min_time) &
    (join_df['time'] <= max_time)
    ]

final_df = join_filter_df \
    .sort_values('time', ascending=True) \
    .set_index('time')
```

Deepnote

Plot the data...

```
final_df.plot.line(subplots=True)
```

```
array([<AxesSubplot:xlabel='time'>, <AxesSubplot:xlabel='time'>],
      dtype=object)
```



Calculate the correlation...

```
final_df['deaths'].corr(final_df['UKgas'])
```

```
0.6094685074564089
```

Deepnote

ON THE MARK DATA

Final Thoughts:

Though this is a very simple analysis, it shows you how easy test data can become available to you via *pydataset*.

How can you use this in your data workflow?
- Mock data for unit tests
- Simple data to test out unfamiliar functions
- Quick data for demos

Link to the Deepnote project with code in the comments!