Using pandas to get started with ALT2

A presentation for LCCS teachers

by Joe English (PDST)

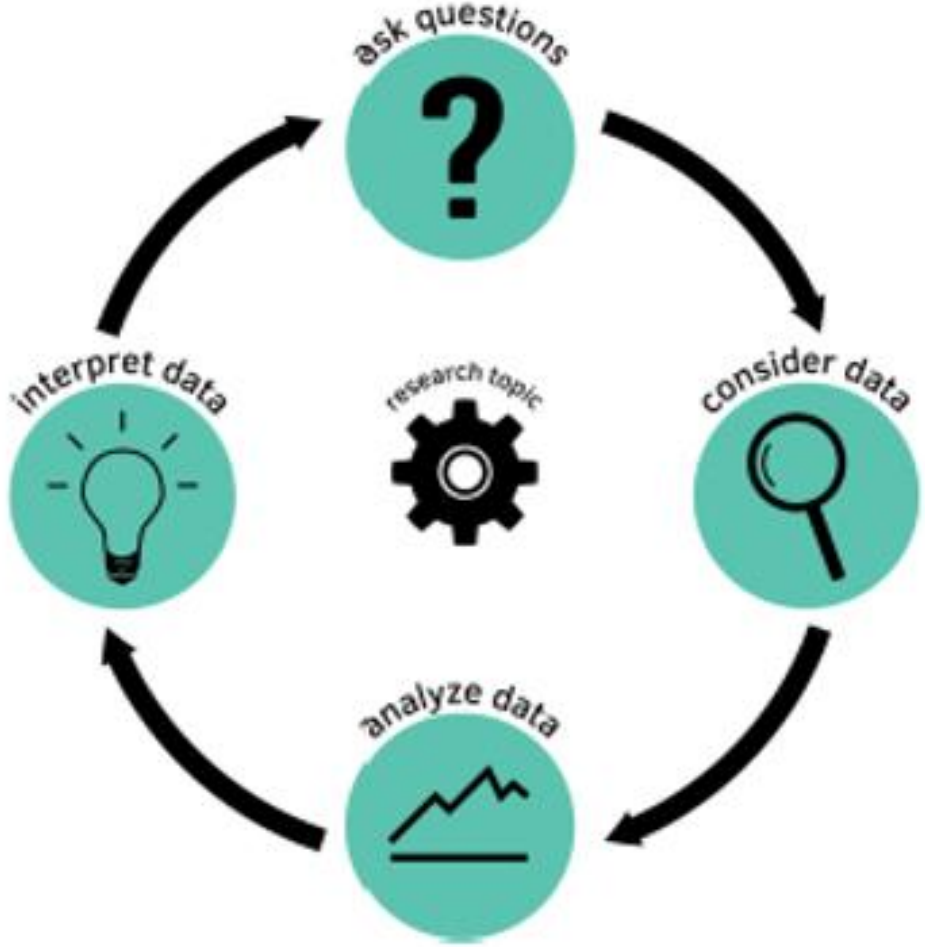
Spring 2021 CoP/Clusters

Applied Learning Task 2

Hypothesis vs. Dataset

The Data Cycle

A framework for ALT2



Data Sets

kaggle

amazon

yelp

Census AtSchool

airbnb

unicef

worldometer

DATA.GOV.IE

An Phríomh-Oifig Staidrimh Central Statistics Office

kaggle

PDST
Professional Development Service for Teachers | An tSeirbhís um Fhorbairt Ghairmiúil do Mhúinteoirí

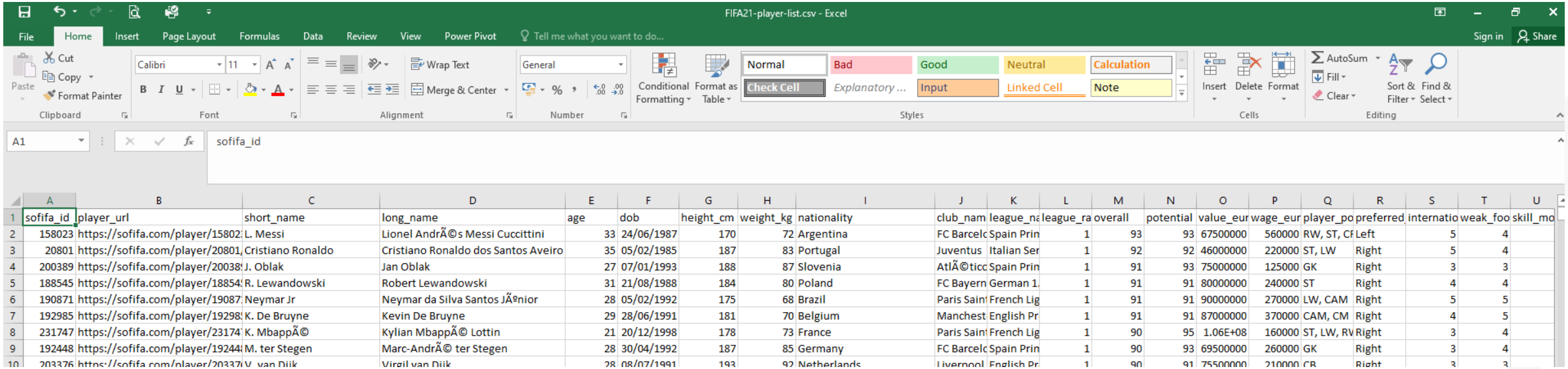
Searchable repository of user-generated datasets (and data challenges)

Detailed and user-friendly search function

Free courses on Python, Machine Learning, Pandas, SQL, etc.

We will explore a FIFA 21 player dataset

Fifa 21 player dataset



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	sofifa_id	player_url	short_name	long_name	age	dob	height_cm	weight_kg	nationality	club_nam	league_na	league_ra	overall	potential	value_eur	wage_eur	player_po	preferred	internatio	weak_foo	skill_mo
2	158023	https://sofifa.com/player/15802	L. Messi	Lionel AndrÃ©s Messi Cuccittini	33	24/06/1987	170	72	Argentina	FC Barcelc	Spain Prin	1	93	93	67500000	560000	RW, ST, CF	Left	5	4	
3	20801	https://sofifa.com/player/20801	Cristiano Ronaldo	Cristiano Ronaldo dos Santos Aveiro	35	05/02/1985	187	83	Portugal	Juventus	Italian Ser	1	92	92	46000000	220000	ST, LW	Right	5	4	
4	200389	https://sofifa.com/player/20038	J. Oblak	Jan Oblak	27	07/01/1993	188	87	Slovenia	AtlÃ©tic	Spain Prin	1	91	93	75000000	125000	GK	Right	3	3	
5	188545	https://sofifa.com/player/18854	R. Lewandowski	Robert Lewandowski	31	21/08/1988	184	80	Poland	FC Bayern	German 1.	1	91	91	80000000	240000	ST	Right	4	4	
6	190871	https://sofifa.com/player/19087	Neymar Jr	Neymar da Silva Santos JÃºnior	28	05/02/1992	175	68	Brazil	Paris Sain	French Lig	1	91	91	90000000	270000	LW, CAM	Right	5	5	
7	192985	https://sofifa.com/player/19298	K. De Bruyne	Kevin De Bruyne	29	28/06/1991	181	70	Belgium	Manchest	English Pr	1	91	91	87000000	370000	CAM, CM	Right	4	5	
8	231747	https://sofifa.com/player/23174	K. MbappÃ©	Kylian MbappÃ© Lottin	21	20/12/1998	178	73	France	Paris Sain	French Lig	1	90	95	1.06E+08	160000	ST, LW, RV	Right	3	4	
9	192448	https://sofifa.com/player/19244	M. ter Stegen	Marc-AndrÃ© ter Stegen	28	30/04/1992	187	85	Germany	FC Barcelc	Spain Prin	1	90	93	69500000	260000	GK	Right	3	4	
10	203376	https://sofifa.com/player/20337	V. van Dijk	Virgil van Dijk	28	08/07/1991	193	92	Netherlands	Liverpool	English Pr	1	90	91	75500000	210000	CB	Right	3	3	
18936	257523	https://sofifa.com/player/25752	Wang Zhen'ao	Zhen'ao Wang	20	10/08/1999	175	69	China PR	Dalian YiF	Chinese S	1	47	57	50000	2000	RW	Right	1	3	
18937	247223	https://sofifa.com/player/24722	Xia Ao	åŤåŸ¥	21	11/02/1999	178	66	China PR	Wuhan Za	Chinese S	1	47	55	40000	1000	CB	Right	1	2	
18938	255626	https://sofifa.com/player/25562	Zhong Jiyu	Jiyu Zhong	23	05/01/1997	170	62	China PR	Shijiazhua	Chinese S	1	47	55	45000	1000	CM	Right	1	2	
18939	257689	https://sofifa.com/player/25768	Wang Huapeng	Huapeng Wang	20	05/08/1999	181	77	China PR	Guangzho	Chinese S	1	47	53	35000	1000	CB	Right	1	2	
18940	257933	https://sofifa.com/player/25793	Huang Wenzhou	Wenzhuo Huang	21	07/01/1999	174	68	China PR	Shanghai	Chinese S	1	47	53	40000	1000	CM	Right	1	2	
18941	256679	https://sofifa.com/player/25667	K. Angulo	Kevin Angulo	24	13/04/1996	176	73	Colombia	AmÃ©ric	Colombia	1	47	52	40000	500	CM	Right	1	2	
18942	257710	https://sofifa.com/player/25771	Zhang Mengxuan	Mengxuan Zhang	21	26/04/1999	177	70	China PR	Chongqin	Chinese S	1	47	52	35000	1000	CB	Right	1	2	
18943	250989	https://sofifa.com/player/25098	Wang Zhenghao	çŽ‹æ”¿è±ª	20	28/06/2000	185	74	China PR	Tianjin TE	Chinese S	1	47	51	35000	1000	CB	Right	1	2	
18944	257697	https://sofifa.com/player/25769	Chen Zitong	Zitong Chen	23	20/02/1997	186	80	China PR	Shijiazhua	Chinese S	1	47	51	40000	1000	CM	Right	1	2	
18945	257936	https://sofifa.com/player/25793	Song Yue	Yue Song	28	20/11/1991	185	79	China PR	Tianjin TE	Chinese S	1	47	47	30000	2000	CM	Right	1	2	

Source: https://www.kaggle.com/stefanoleone992/fifa-21-complete-player-dataset

Pandas

Many popular Python libraries:

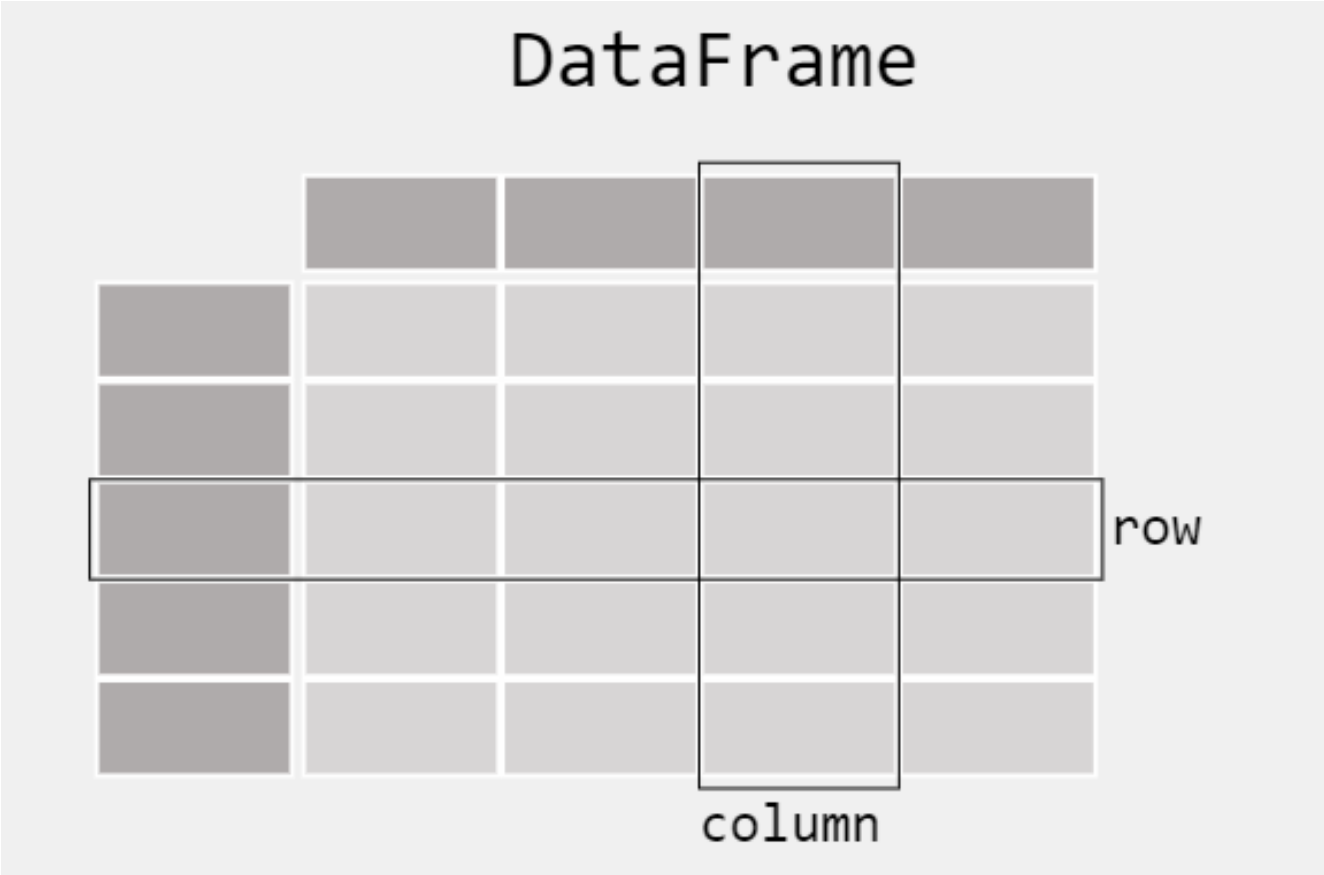
- NumPy http://www.numpy.org/
- SciPy https://www.scipy.org/scipylib/
- Pandas http://pandas.pydata.org/
- SciKit-Learn http://scikit-learn.org/

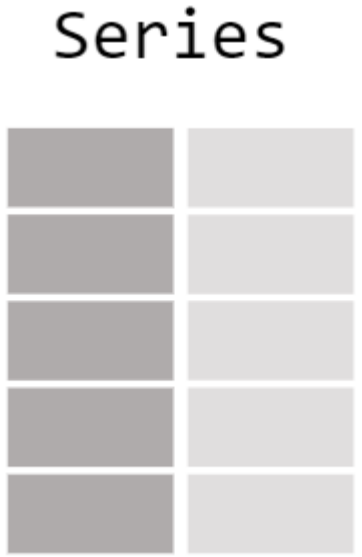
Visualisation libraries

- matplotlib https://matplotlib.org/
- Seaborn https://seaborn.pydata.org/

pandas will help you to explore, clean and process your data

What kind of data does pandas handle?





Each column is called a **Series**

In pandas, a data table is called a **DataFrame**

PDST

Pandas – reading in a CSV file

```
# Using pandas - recommended for larger files
import pandas

# Read the entire CSV file into a pandas DataFrame
fifa_df = pandas.read_csv('FIFA21-player-list.csv')

# Display the length of the dataframe
print("Nr. rows", len(fifa_df))

# Display the number of rows and column
print("Shape (rows, cols)", fifa_df.shape)
```

Output looks like this:

```
Nr. rows 18944
Shape (rows, cols) (18944, 106)
```

Pandas – selecting specific columns from a DataFrame

	short_name	age	dob	height_cm	weight_kg	nationalit	club_nam	value_eur	wage_eur	player_po	preferred_
2	L. Messi	33	24/06/1987	170	72	Argentina	FC Barcelo	67500000	560000	RW, ST, CF	Left
3	Cristiano Ronaldo	35	05/02/1985	187	83	Portugal	Juventus	46000000	220000	ST, LW	Right
4	J. Oblak	27	07/01/1993	188	87	Slovenia	AtlÃ©tico	75000000	125000	GK	Right
5	R. Lewandowski	31	21/08/1988	184	80	Poland	FC Bayern	80000000	240000	ST	Right
6	Neymar Jr	28	05/02/1992	175	68	Brazil	Paris Sain	90000000	270000	LW, CAM	Right
7	K. De Bruyne	29	28/06/1991	181	70	Belgium	Manchest	87000000	370000	CAM, CM	Right

We may only be interested in the player's name, age and value

Pandas

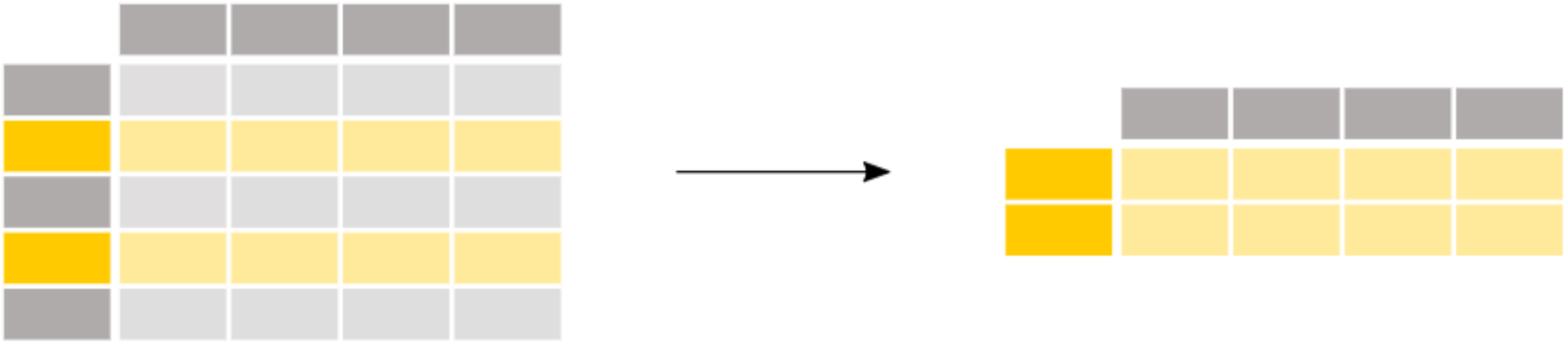
```
# Using pandas - recommended for larger files
import pandas

# Read the entire CSV file into a pandas DataFrame
fifa_df = pandas.read_csv('FIFA21-player-list.csv')

# Select a number of columns - all rows
fifa_df1 = fifa_df[['short_name', 'age', 'value_eur']]
print(fifa_df1) # DataFrame
```

Output looks like this:

	short_name	age	value_eur
0	L. Messi	33	67500000
1	Cristiano Ronaldo	35	46000000
2	J. Oblak	27	75000000
3	R. Lewandowski	31	80000000
4	Neymar Jr	28	90000000
...	...	...	...
18939	K. Angulo	24	40000
18940	Zhang Mengxuan	21	35000
18941	Wang Zhenghao	20	35000
18942	Chen Zitong	23	40000
18943	Song Yue	28	30000

```
# Select a single column - all rows
values = fifa_df['value_eur']
```

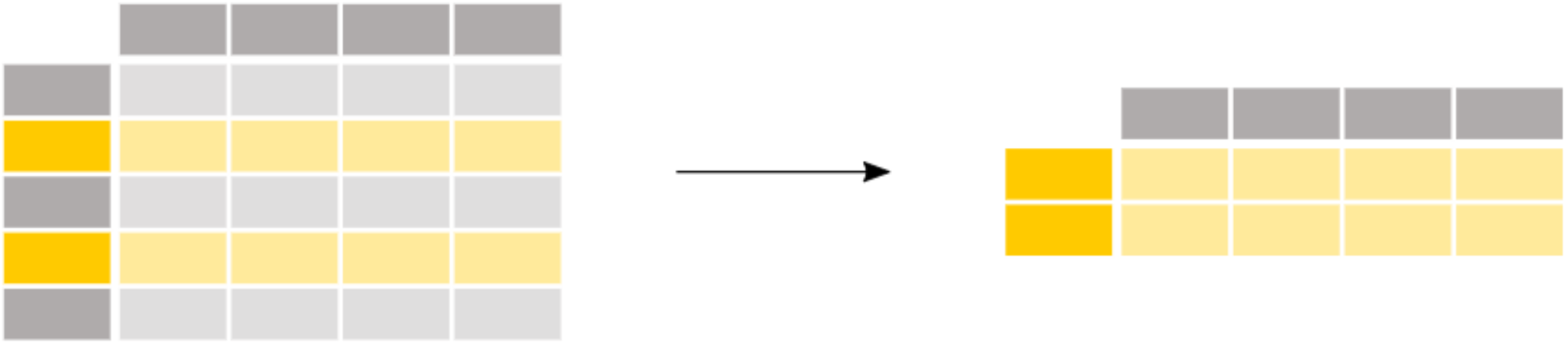
Pandas – selecting specific rows from a DataFrame



1	short_name	age	dob	height_cm	weight_kg	nationalit	club_nam	value_eur	wage_eur	player_po	preferred_
2	L. Messi	33	24/06/1987	170	72	Argentina	FC Barcelc	67500000	560000	RW, ST, CF	Left
3	Cristiano Ronaldo	35	05/02/1985	187	83	Portugal	Juventus	46000000	220000	ST, LW	Right
4	J. Oblak	27	07/01/1993	188	87	Slovenia	AtlÃ©ticc	75000000	125000	GK	Right
5	R. Lewandowski	31	21/08/1988	184	80	Poland	FC Bayern	80000000	240000	ST	Right
6	Neymar Jr	28	05/02/1992	175	68	Brazil	Paris Sain	90000000	270000	LW, CAM	Right
7	K. De Bruyne	29	28/06/1991	181	70	Belgium	Manchest	87000000	370000	CAM, CM	Right

We may only be interested in players that meet certain criteria

Pandas – selecting specific rows from a DataFrame



1	short_name	age	dob	height_cm	weight_kg	nationalit	club_nam	value_eur	wage_eur	player_po	preferred_
2	L. Messi	33	24/06/1987	170	72	Argentina	FC Barcelc	67500000	560000	RW, ST, CF	Left
3	Cristiano Ronaldo	35	05/02/1985	187	83	Portugal	Juventus	46000000	220000	ST, LW	Right
4	J. Oblak	27	07/01/1993	188	87	Slovenia	AtlÃ©ticc	75000000	125000	GK	Right
5	R. Lewandowski	31	21/08/1988	184	80	Poland	FC Bayern	80000000	240000	ST	Right
6	Neymar Jr	28	05/02/1992	175	68	Brazil	Paris Sain	90000000	270000	LW, CAM	Right
7	K. De Bruyne	29	28/06/1991	181	70	Belgium	Manchest	87000000	370000	CAM, CM	Right

We may only be interested in players that meet certain criteria

Filter by nationality

```
irish_players = fifa_df1[fifa_df1['nationality'] == 'Republic of Ireland']
print(irish_players)
```

Output looks like this:

	short_name	age	...	value_eur	wage_eur
347	M. Doherty	28	...	15000000	96000
614	S. Coleman	31	...	9000000	76000
753	J. Egan	27	...	11000000	36000
826	E. Stevens	29	...	8500000	37000
994	S. Duffy	28	...	8500000	50000
...	...	...	...	...	...
18911	R. Dinanga	18	...	45000	500
18912	J. Browne	19	...	45000	500
18913	P. McGarvey	16	...	30000	500
18916	A. Cetiner	18	...	40000	500
18918	A. Phelan	19	...	40000	500

[338 rows x 7 columns]

Any Boolean operator can be used to subset the data:

> greater; >= greater or equal;
< less; <= less or equal;
== equal; != not equal;

Premier league players over 30

```
prem_players_over_30 = fifa_df1[(fifa_df1['league_name'] == 'English Premier League') & (fifa_df1['age'] > 30)]
print(prem_players_over_30)
```

Output looks like this:

	short_name	age	...	value_eur	wage_eur
13	S. Agüero	32	...	53000000	300000
36	H. Lloris	33	...	27000000	125000
41	P. Aubameyang	31	...	45000000	170000
61	J. Vardy	33	...	28000000	160000
82	Thiago Silva	35	...	11500000	93000
...	...	...	...	...	...
5959	L. Peltier	33	...	475000	30000
6947	D. Button	31	...	575000	21000
9127	E. Jakupović	35	...	170000	9000
9131	D. Martin	34	...	275000	8000
10341	S. Henderson	32	...	325000	9000

[81 rows x 7 columns]

Can combine Boolean expressions

& and
| or
~ not

Pandas – selecting specific rows and columns from a DataFrame

1	short_name	age	dob	height_cr	weight_kg	nationalit	club_nam	value_eur	wage_eur	player_po	preferred
2	L. Messi	33	24/06/1987	170	72	Argentina	FC Barcelc	67500000	560000	RW, ST, CF	Left
3	Cristiano Ronaldo	35	05/02/1985	187	83	Portugal	Juventus	46000000	220000	ST, LW	Right
4	J. Oblak	27	07/01/1993	188	87	Slovenia	AtlÃ©ticc	75000000	125000	GK	Right
5	R. Lewandowski	31	21/08/1988	184	80	Poland	FC Bayern	80000000	240000	ST	Right
6	Neymar Jr	28	05/02/1992	175	68	Brazil	Paris Sain	90000000	270000	LW, CAM	Right
7	K. De Bruyne	29	28/06/1991	181	70	Belgium	Manchest	87000000	370000	CAM, CM	Right

We may only be interested in seeing the player names of players over 30

Filter by row and column

```
players_over_30 = fifa_df1.loc[(fifa_df1['age'] > 30), ['short_name', 'age']]
print(players_over_30)
```

Output looks like this:

	short_name	age
0	L. Messi	33
1	Cristiano Ronaldo	35
3	R. Lewandowski	31
13	S. Agüero	32
14	Sergio Ramos	34
...	...	...
18302	Song Zhiwei	31
18507	J. Russell	35
18508	G. Maley	37
18884	Gao Xiang	31
18927	Wang Jianwen	32

[2803 rows x 2 columns]

The `loc` operator is required in front of the selection brackets, [].

When using `loc` the part before the comma is the rows you want, and the part after the comma is the columns you want to select.

In this example we want:

- all the rows where the *age* is greater than 30 and
- all the *short_name* and *age* columns

Aggregrating Data

Available methods:

```
value_counts()
max(), min()
mean(), median()
describe()
```

Plus more …

Count players by nationality

```
fifa_df1 = fifa_df[['short_name', 'age', 'nationality', 'club_name', 'league_name', 'value_eur', 'wage_eur']]

nationality_counts = fifa_df1['nationality'].value_counts()
print(nationality_counts)
```

Output looks like this:

```
England          1685
Germany          1189
Spain            1072
France            984
Argentina         936
                  ...
Malaysia            1
Saint Lucia         1
Indonesia           1
Aruba               1
Andorra             1
Name: nationality, Length: 162, dtype: int64
```

`value_counts` counts the number of entries in each category of a variable

Minumum, Maximum and Averages

```
fifa_df1 = fifa_df[['short_name', 'age', 'nationality', 'club_name', 'league_name', 'value_eur', 'wage_eur']]

# What is the average player age?
print("Average player age:", fifa_df1['age'].mean())

# What are the youngest and oldest ages
print("Youngest:", fifa_df1['age'].min())
print("Oldest:", fifa_df1['age'].max())

# What is the median age and wage of a player?
print("Median age/wage:\n", fifa_df1[['age', 'wage_eur']].median())
```

Output looks like this:

```
Average player age: 25.22582347972973
Youngest: 16
Oldest: 53
Median age/wage:
 age           25.0
wage_eur    3000.0
dtype: float64
```

Grouping Data

Let's say we wanted to see the average wages by player age


```
fifa_df1 = fifa_df[['short_name', 'age', 'nationality', 'club_name', 'league_name', 'value_eur', 'wage_eur']]

# Player wage grouped by age?
print(fifa_df1.groupby('age')['wage_eur'].mean())
```

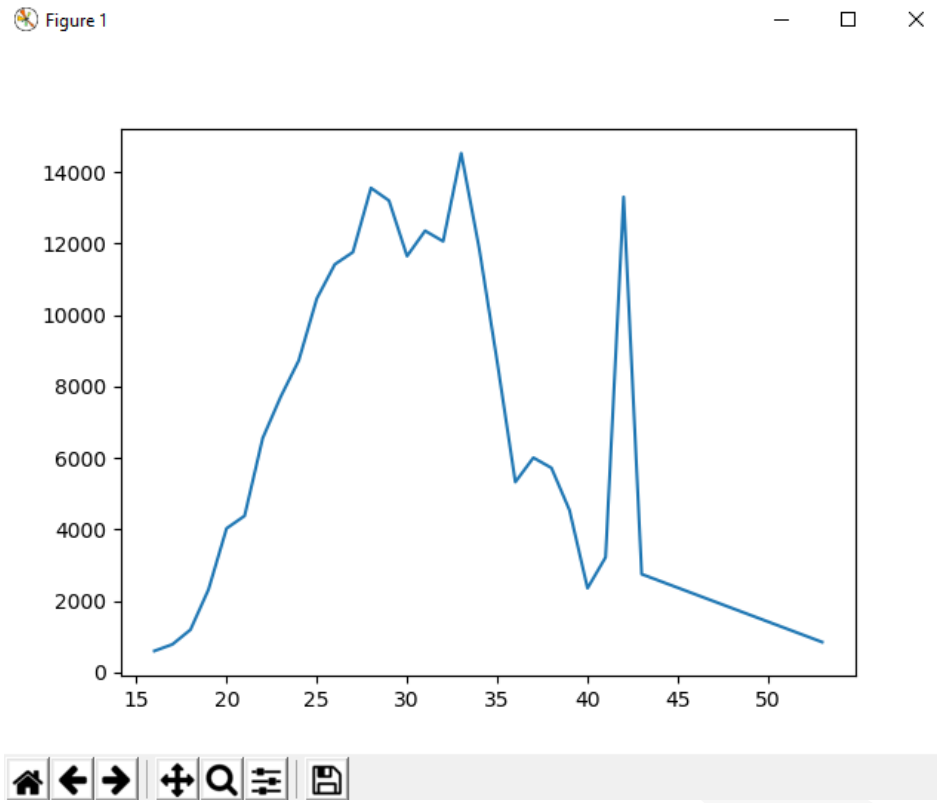
Grouping Data

Output

age	wage_eur
16	601.724138
17	785.897436
18	1196.257485
19	2322.093023
20	4029.747722
21	4380.477223
22	6550.801394
23	7712.287848
24	8726.177091
25	10458.400000
26	11415.700861
27	11755.464256
28	13551.255887
29	13197.990431
30	11645.508021
31	12352.929688
…	
43	2750.000000
53	850.000000

Plotting grouped data ...

```
from matplotlib import pyplot

# Player wage grouped by age
df = fifa_df1.groupby('age')['wage_eur'].mean()
print(df)
pyplot.plot(df)
pyplot.show()
```

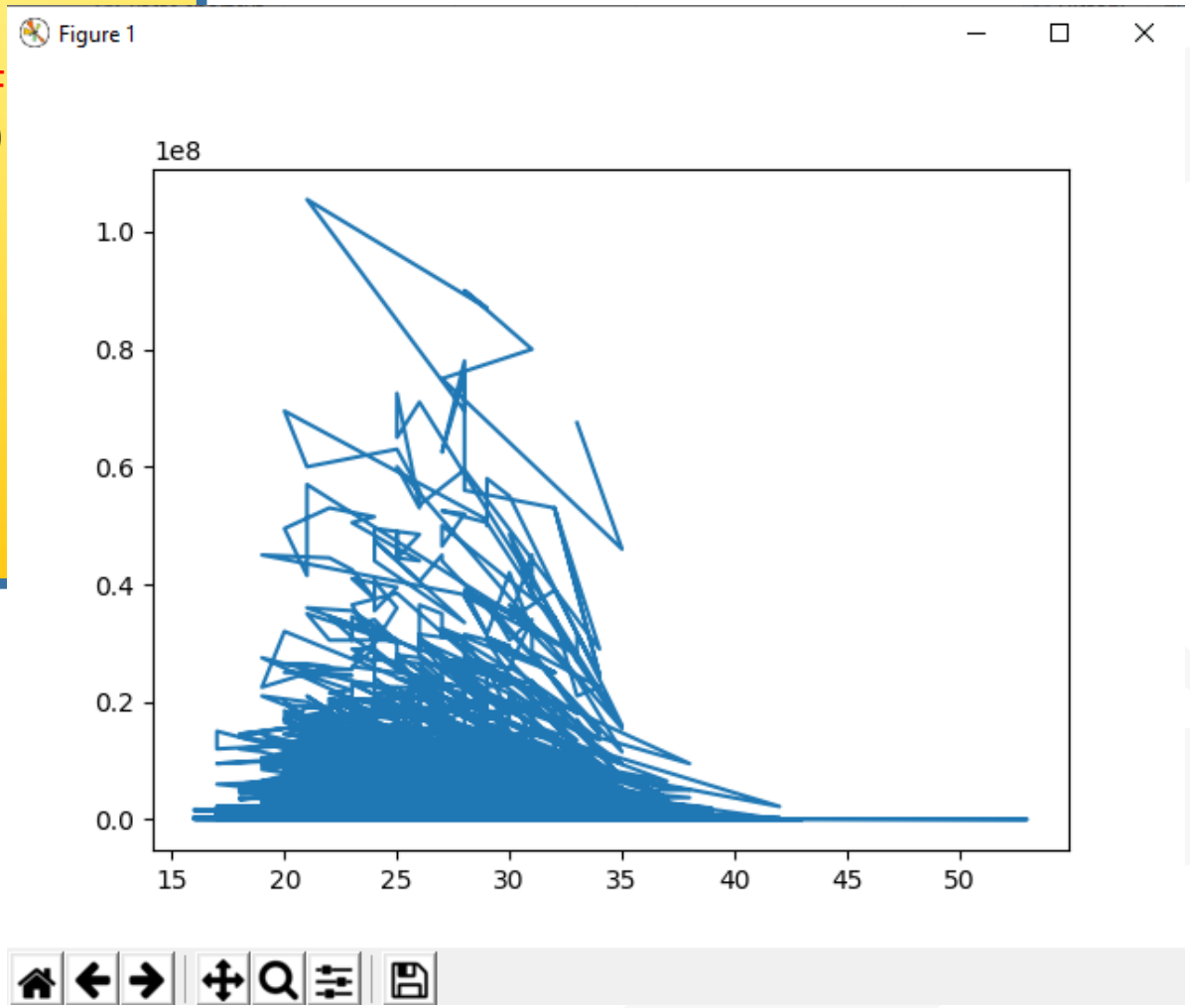


Output

age	wage_eur
16	601.724138
17	785.897436
18	1196.257485
19	2322.093023
20	4029.747722
21	4380.477223
22	6550.801394
23	7712.287848
24	8726.177091
25	10458.400000
26	11415.700861
27	11755.464256
28	13551.255887
29	13197.990431
30	11645.508021
31	12352.929688
…	
43	2750.000000
53	850.000000

Attempt #1

```
# Using pandas - recommended for larger files
import pandas
from matplotlib import pyplot

# Read the entire CSV file into a pandas DataF
df = pandas.read_csv('FIFA21-player-list.csv')

# Filter out the column, value_eur
player_values = df['value_eur']
player_ages = df['age']

pyplot.plot(player_ages, player_values)
pyplot.show()
```



Player wage by league

```
print(fifa_df1.groupby(['league_name'])['wage_eur'].mean())
```

Output looks like this:

league_name	
Argentina Primera División	6421.547800
Argentinian Primera B Nacional	500.000000
Australian Hyundai A-League	1594.000000
Austrian Football Bundesliga	5522.258065
Belgian Jupiler Pro League	6365.650407
Campeonato Brasileiro Série A	9341.666667
Chilian Campeonato Nacional	1020.402299
Chinese Super League	5294.724771
Colombian Liga Postobón	1007.179487
Croatian Prva HNL	500.000000
Czech Republic Gambrinus Liga	500.000000
Danish Superliga	4457.704918
Ecuadorian Serie A	500.000000
English League Championship	9866.361072
English League One	2133.571429
English League Two	2207.024793
English Premier League	52149.082569
Finnish Veikkausliiga	500.000000
French Ligue 1	19445.583333
French Ligue 2	2111.080586
German 1. Bundesliga	24073.996350
German 2. Bundesliga	5720.611440
German 3. Bundesliga	1285.153257
Greek Super League	500.000000
Holland Eredivisie	4582.618026
Italian Serie A	27602.945736
Italian Serie B	500.000000
Japanese J. League Division 1	2773.416507

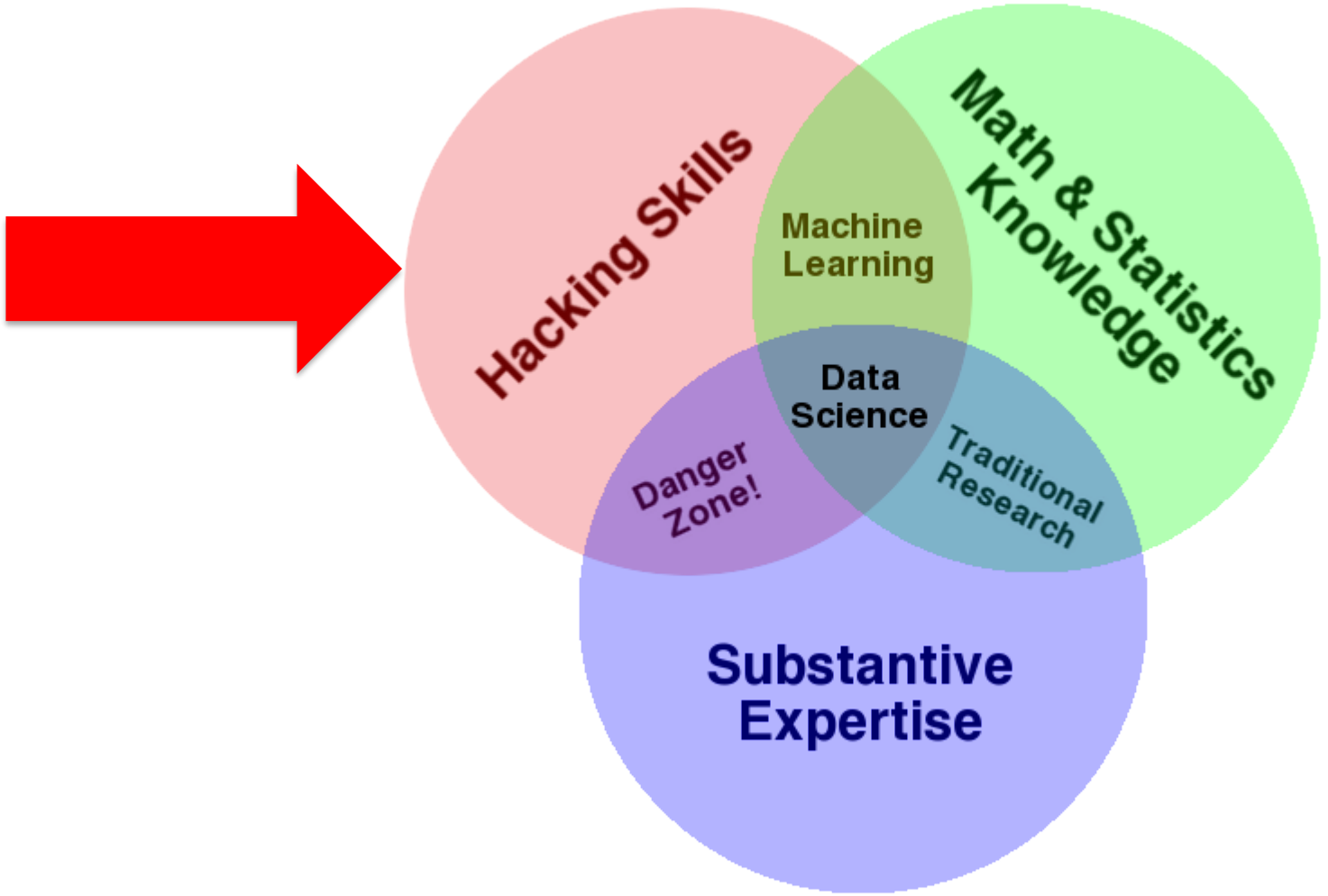
Player wage by nationality and age

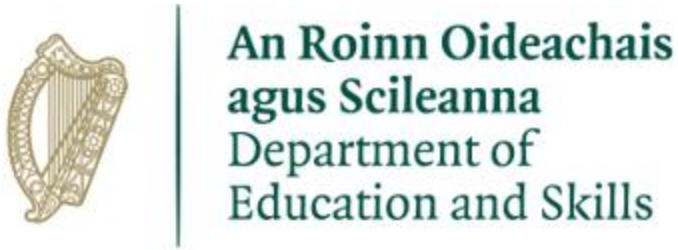
```
print(fifa_df1.groupby(['nationality', 'age'])['wage_eur'].mean().head(20))
```

Output looks like this:

```
nationality  age
Afghanistan  22      2000.000000
             28      1000.000000
Albania  18       900.000000
         19      1000.000000
         20      5371.428571
         21      1500.000000
         22      2000.000000
         23     16625.000000
         24      1950.000000
         25     30000.000000
         26     15250.000000
         27     22500.000000
         28      5750.000000
         29      8200.000000
         30      7333.333333
         31       500.000000
         32       950.000000
         34      4000.000000
Algeria      20      9000.000000
         22     13500.000000
Name: wage_eur, dtype: float64
```

Data Science – One Definition




An Roinn Oideachais
agus Scileanna
Department of
Education and Skills


BY
SA