Data Ranking & Filtering

Ranking and Sorting

Sorting options- sort_values

Sorting by 1 column

```
salaries.sort_values("base_salary")
```

Sorting by a few columns, must use a subset []

```
salaries.sort_values(["club","base_salary"] )
```

Using inplace to make the changes permanent

```
salaries.sort_values("club", inplace=True)
```

Sorting options- sort_index

Sorting by index (restoring the default order)

```
salaries.sort_index()
```

Sorting by reversed index

```
salaries.sort_index(ascending=False, inplace=True)
```

Ranking

To use the rank() method we will need to do some cleaning to remove all NaN values or the rank() method won't wo
By default the ranking will be according to ascending order
By default rank() will use float data type

```
salaries["base_salary"].rank()
```

Ranking

Changing data type

```
salaries["base_salary"].rank().astype("int")
```

Changing to descnding order

```
salaries.sort_values("base_salary", ascending=False)
```

Data Filtering

Creating a filter

Boolean operators:

>,<, >=,<=,==, !=

```
# # broadcasting boolean operation: use == and not = ⚠️
df["team"] == "engineering"
```

```
0      False
1       True
2       True
3      False
4      False
       ...
995    False
996    False
997    False
998     True
999    False
Name: team, Length: 1000, dtype: bool
```

Creating a filter

```
df["team"] == "engineering"
```

```
0      False
1       True
2       True
3      False
4      False
       ...
995    False
996    False
997    False
998     True
999    False
Name: team, Leng
```

```
# wrap df[] around the conditional operation to extract the values of the condition
df[df["team"] == "engineering"]
```

	first_name	last_name	salary	start_date	gender	remote	team
1	Coretta	McEvon	637457	3/20/2020	Male	False	engineering
2	Clarette	Tarbett	977749	11/22/2020	Agender	True	engineering
5	Auberta	Whistlecraft	510781	NaN	Polygender	False	engineering
6	Devland	Cominetti	194815	1/16/2021	Bigender	False	engineering
9	Susanna	Ivachyov	873134	11/19/2020	Agender	False	engineering
...	...	...	...	...	...	...	...
972	Kennett	Franzonello	580290	7/7/2020	Non-binary	True	engineering
973	Ola	Dautry	804650	8/20/2020	Agender	True	engineering
977	Theda	Sharpe	739171	5/29/2020	Bigender	True	engineering
988	Ariana	Culverhouse	901497	12/4/2020	Female	False	engineering
998	Jerald	Penella	862555	9/8/2020	Female	False	engineering

230 rows × 7 columns

Note, only rows that has “True” value returned, due to it being many rows we don’t see all of them in the slide

Using a mask to filter

```
df["team"] == "engineering"
```

```
# instead of wrapping df[] around the conditional operation, use a variable name
female_mask = df["gender"] == "Female"
df[female_mask]
```

Combining filters with and/or

```
df[mask1 | mask2]
```

OR

```
df[mask1 & mask2]
```

AND

Special methods operators:

.isin()

```
mask = df["team"].isin(["data analytics","management","engineering"])
```

.isnull()

```
mask = df['gender'].isnull()
```