

Datawhys Internship Bootcamp

Discussion session

Filtering

Discussion:

- What does reading a dataset mean?
- Where do we store our read data?
- How do we view our stored data?
- Why do we have to use filtering in a dataset?
- Why sometimes we create a new column in a dataset?
- What is the purpose of the 'value_counts' function?

Discussion with Example:

The Dataset : “trees2.csv”

Q. What do we do first with this “trees2.csv” dataset?

- Import the **pandas** library
- **Read** the dataset using pandas

Q. Where do we store the data we have now?

- Store the data in a **variable** (dataframe)

Q. How do we see the data we stored in the variable?

- Using the variable name **Treedata**

Treedata =

	Girth	Height	Volume	Type
0	8.3	70.0	10.3	cherry
1	8.6	65.0	10.3	cherry
2	8.8	63.0	10.2	cherry
3	10.5	72.0	16.4	cherry
4	10.7	81.0	18.8	cherry
...
57	17.9	69.2	47.1	plum
58	18.8	67.3	50.5	plum
59	19.7	67.4	55.6	plum
60	17.2	67.3	42.3	plum
61	21.1	73.8	69.8	plum

Objective:

To find the feature(s) that distinguish the tree types (cherry or plum)

Using filtering in Treedata :

- To divide the **Treedata** by tree “Type”
- To compare the two species by using the features
- To find the feature(s) that distinguishes the tree species

Find distinction between the species :

- Filter the data by tree type

Cherrydata =

	Girth	Height	Volume	Type
0	8.3	70.0	10.3	cherry
1	8.6	65.0	10.3	cherry
2	8.8	63.0	10.2	cherry
3	10.5	72.0	16.4	cherry
4	10.7	81.0	18.8	cherry
5	10.8	83.0	19.7	cherry
6	11.0	66.0	15.6	cherry
7	11.0	75.0	18.2	cherry
8	11.1	80.0	22.6	cherry
9	11.2	75.0	19.9	cherry
10	11.3	79.0	24.2	cherry
11	11.4	76.0	21.0	cherry
12	11.4	76.0	21.4	cherry
13	11.7	69.0	21.3	cherry
14	12.0	75.0	19.1	cherry

Plumdata =

	Girth	Height	Volume	Type
31	9.2	57.4	10.3	plum
32	8.6	52.3	8.2	plum
33	8.4	50.3	7.5	plum
34	9.6	58.7	11.5	plum
35	10.1	68.3	14.8	plum
36	12.4	69.8	22.8	plum
37	10.8	53.0	13.1	plum
38	10.1	61.7	13.4	plum
39	9.6	66.8	13.1	plum
40	9.4	62.0	11.6	plum
41	11.4	65.9	18.2	plum
42	11.9	63.5	19.1	plum
43	11.0	62.5	16.1	plum
44	10.0	56.3	12.0	plum
45	12.6	62.4	21.0	plum

Which feature is decisive to make distinction between the tree types:

Calculate `mean()` of the features of each tree type

Compare the corresponding features with each tree type

Check which features has the biggest difference from each other

Finally, the features which has the biggest difference will be the defining factor in making distinction of the trees.

Calculate the mean of all features:

Cherrydata =

	Girth	Height	Volume	Type
0	8.3	70.0	10.3	cherry
1	8.6	65.0	10.3	cherry
2	8.8	63.0	10.2	cherry
3	10.5	72.0	16.4	cherry
4	10.7	81.0	18.8	cherry
5	10.8	83.0	19.7	cherry
6	11.0	66.0	15.6	cherry
7	11.0	75.0	18.2	cherry
8	11.1	80.0	22.6	cherry
9	11.2	75.0	19.9	cherry
10	11.3	79.0	24.2	cherry
11	11.4	76.0	21.0	cherry
12	11.4	76.0	21.4	cherry
13	11.7	69.0	21.3	cherry
14	12.0	75.0	19.1	cherry

Cherrydata['Girth'].
mean()

Cherrydata['Height'].
mean()

Cherrydata['Volume']
.mean()

Calculate the mean of all features:

Plumdata =

	Girth	Height	Volume	Type
31	9.2	57.4	10.3	plum
32	8.6	52.3	8.2	plum
33	8.4	50.3	7.5	plum
34	9.6	58.7	11.5	plum
35	10.1	68.3	14.8	plum
36	12.4	69.8	22.8	plum
37	10.8	53.0	13.1	plum
38	10.1	61.7	13.4	plum
39	9.6	66.8	13.1	plum
40	9.4	62.0	11.6	plum
41	11.4	65.9	18.2	plum
42	11.9	63.5	19.1	plum
43	11.0	62.5	16.1	plum
44	10.0	56.3	12.0	plum
45	12.6	62.4	21.0	plum

Plumdata['Girth'].
mean()

Plumdata['Height'].m
ean()

Plumdata['Volume'].
mean()

Compare the feature values for each tree type:

	mean(Girth)	mean(Height)	mean(Volume)
Cherry:	13.25	76.0	30.18
Plumdata:	13.12	63.08	25.47

Which feature has the biggest difference:

	mean(Girth)	mean(Height)	mean(Volume)
Cherry:	13.25	76.0	30.18
Plumdata:	13.12	63.08	25.47

Final Result:

So the feature that defines if a tree is plum or cherry is their **Height**.