Microsoft: DAT210x Programming with Python for Data Science

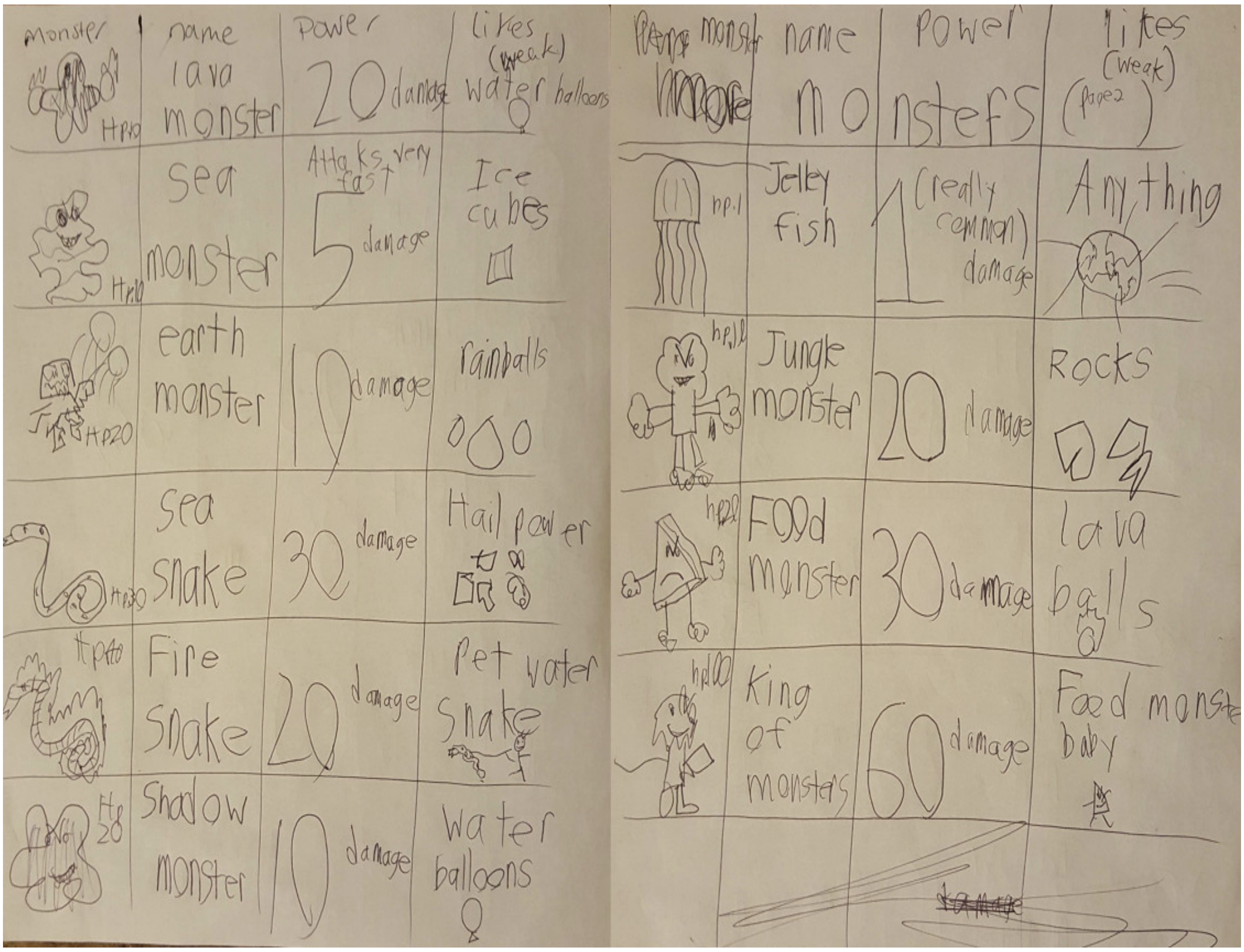
Help

Course › 5. Data Modeling › Lecture: Clustering › Clustering

Clustering

Bookmark this page

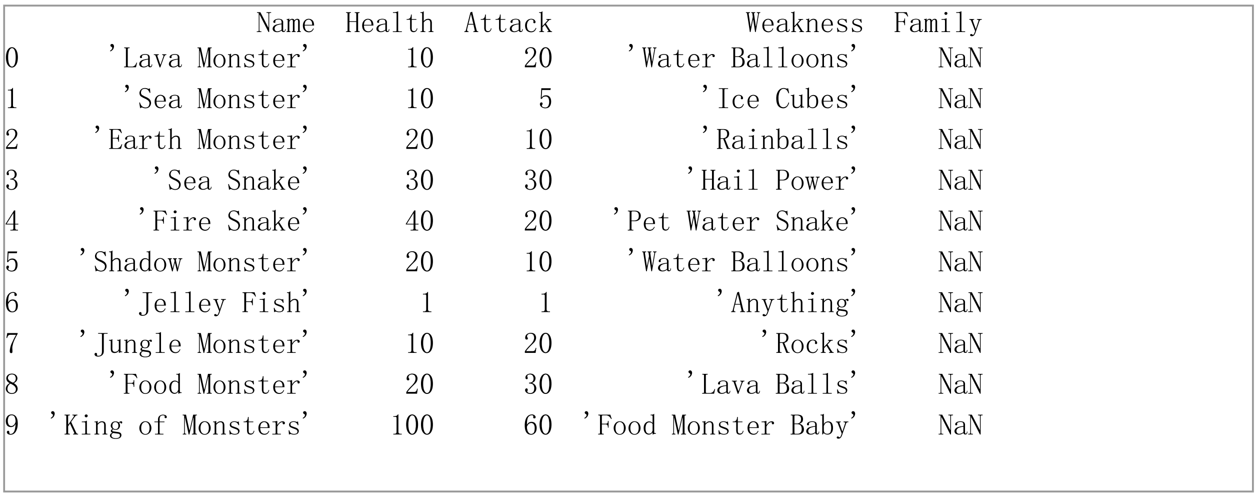
My nephew wants me to program a video game for him called 'Monster Family', and even provided a description of its characters:



Monster	name	Power	likes (weak)
HP10	lava monster	20 damage	water balloons
HP10	sea monster	Attacks very fast 5 damage	Ice cubes
HP20	earth monster	10 damage	rainballs
HP30	Sea Snake	30 damage	Hail power
HP40	Fire Snake	20 damage	Pet water Snake
HP 20	Shadow monster	10 damage	water balloons

monster	name	Power	likes (weak)
More	monsters		(page2)
hp.1	Jelley fish	1 (really common) damage	Anything
hp.10	Jungle monster	20 damage	Rocks
hp20	Food monster	30 damage	lava balls
hp100	King of monsters	60 damage	Food monster baby
			~~damage~~

Although he's young, you can tell he's particular about details. To make the game to his likings, I needed to know which monsters were real family members, and which monsters weren't. If he had added just one more column that held that detail, FamilyA, FamilyB, etc., I would have been set:

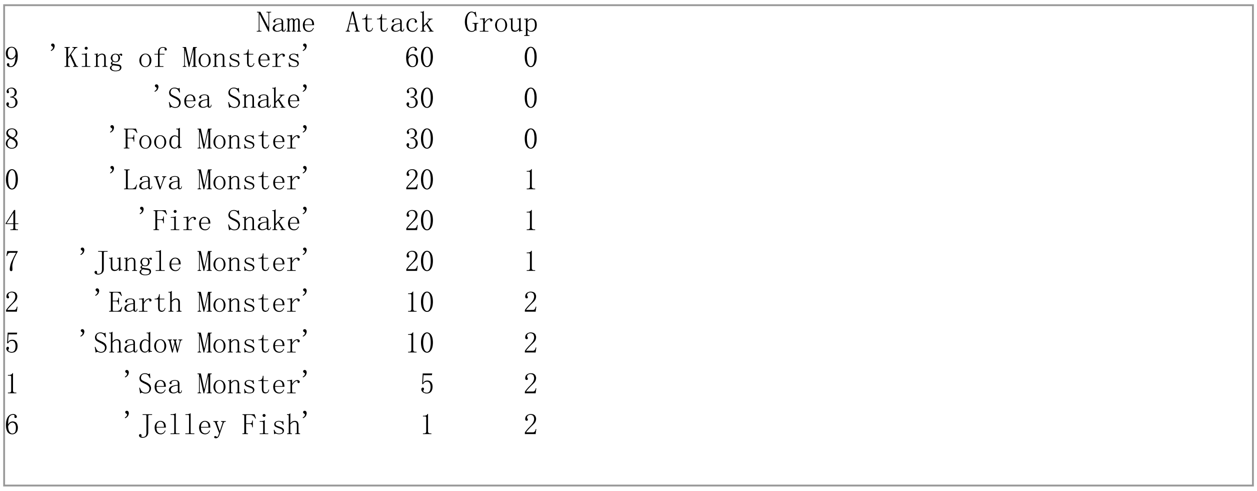


```
                 Name  Health  Attack            Weakness  Family
0       'Lava Monster'      10      20     'Water Balloons'     NaN
1        'Sea Monster'      10       5         'Ice Cubes'     NaN
2      'Earth Monster'      20      10         'Rainballs'     NaN
3          'Sea Snake'      30      30        'Hail Power'     NaN
4         'Fire Snake'      40      20   'Pet Water Snake'     NaN
5     'Shadow Monster'      20      10    'Water Balloons'     NaN
6        'Jelley Fish'       1       1          'Anything'     NaN
7     'Jungle Monster'      10      20             'Rocks'     NaN
8       'Food Monster'      20      30        'Lava Balls'     NaN
9   'King of Monsters'     100      60 'Food Monster Baby'     NaN
```

This monster 'dataset' is similar to real-world data in that it comes loaded with observational features, but isn't *labeled*. The one question I want a direct answer to isn't included as a feature. If there were way to automatically group similar samples based solely on their features, we then could use that knowledge to guide us towards actionable intelligence. That way exist, and its called unsupervised clustering.

Similarity

Since the goal of clustering is the grouping of similar records, you have to first define what similarity means. How would you go define monster similarity?



```
                 Name  Attack  Group
9   'King of Monsters'      60      0
3          'Sea Snake'      30      0
8       'Food Monster'      30      0
0       'Lava Monster'      20      1
4         'Fire Snake'      20      1
7     'Jungle Monster'      20      1
2      'Earth Monster'      10      2
5     'Shadow Monster'      10      2
1        'Sea Monster'       5      2
6        'Jelley Fish'       1      2
```

One way you could group them is by attack power. Perhaps the monsters who deal the most damage to the player belong to a family, the weaker monsters belong to a family, and then the rest grouped as a family too, as shown above.

```
                Name              Weakness  Group
0       'Lava Monster'     'Water Balloons'      0
1        'Sea Monster'          'Ice Cubes'      0
2      'Earth Monster'          'Rainballs'      0
3          'Sea Snake'         'Hail Power'      0
4         'Fire Snake'    'Pet Water Snake'      0
5     'Shadow Monster'     'Water Balloons'      0
6         'Jelley Fish'          'Anything'      1
7     'Jungle Monster'              'Rocks'      2
8       'Food Monster'         'Lava Balls'      3
9   'King of Monsters'  'Food Monster Baby'      4
```

Another grouping would be by weakness. It seems a lot of monsters that share a water-based weakness. All of these might belong to the same family. The remaining monsters might each belong to a separate family, or might actually be members of a sporadic family. There are many other ways you could group them as well, such as by size, by name (e.g. monster vs snake), etc.

Without a generalizable way to group the samples, deterministic computers can't cluster your data. What's needed is a systematic means of measuring the **overall** similarity between your samples. Let's discuss how that's accomplished in the next section.