

DOG BREEDS

DATA SCIENCE PROJECT

By Niv Huga & Yarden Gispan

What's a dog breed?

People have been breeding dogs since prehistoric times. The earliest dog breeders used wolves to create domestic dogs. From the beginning, humans purposefully bred dogs to perform various tasks. Hunting, guarding, and herding are thought to be among the earliest jobs eagerly performed by the animal destined to be called “man’s best friend.”

For thousands of years, humans bred dogs toward the physical and mental traits best suited for the work expected of them. The sleek Greyhound types bred to chase fleet-footed prey, and the huge mastiff types used as guard dogs and warriors, are two ancient examples of dogs bred for specific jobs.



MAIN PROCESS STEPS OF THE PROJECT

MAIN STEPS

OBTAINING DATA

Web crawling
for collecting
& obtaining
the data

SCRUBBING DATA

Cleaning,
formatting and
filtering the data.
Filled the missing
data with API.

EXPLORING DATA

Visualizing and
understanding the data

MODELING DATA

Clustering the
data into groups
and modeling.

INTERPRETING DATA

Presentation of data,
understanding and
delivering the results

WEB CRAWLING FOR COLLECTING & OBTAINING THE DATA

In this step, we will collect information regarding dog breeds.

We will crawl along web pages and scrape information about dog breeds.

Source data: <https://dogtime.com/dog-breeds>

Main tool: BeautifulSoup

All Dog Breed Profiles

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Afador
Affenhuahua
Affenpinscher
Afghan Hound
Airedale Terrier
Akash

Aussiedoodle

The Aussiedoodle is a mixed breed dog — a cross between the Australian Shepherd and Poodle dog breeds. Incredibly smart, playful, and loyal, these pups inherited some of the best qualities from both of their parents.

Aussiedoodles go by several names, including Aussiepoo and Aussiepoodle. Despite their unfortunate status as a designer breed, you may find these mixed breed dogs in shelters and rescues. So please remember to adopt! Don't shop!

These active dogs, often referred to as an "Einstein" breed for their smarts, do well in homes that can provide plenty of attention and exercise. The Aussiedoodle makes an excellent family dog, as long as smaller children know how to safely play with the pups. They are also incredible therapy dogs, given how quickly they bond to a specific human or two.

See below for all mixed dog breed traits and facts about Aussiedoodles!

Aussiedoodle Mixed Dog Breed Pictures

Breed Characteristics:

Adaptability	Friendliness
[+] Adapts Well To Apartment Living	★★★★★
[+] Good For Novice Owners	★★★★★
[+] Sensitivity Level	★★★★★
[+] Tolerates Being Alone	★★★★★
[+] Tolerates Cold Weather	★★★★★
[+] Tolerates Hot Weather	★★★★★

All Around Friendliness

Friendliness	Playfulness
[+] Affectionate With Family	★★★★★
[+] Kid-Friendly	★★★★★
[+] Dog-Friendly	★★★★★

WEB CRAWLING FOR COLLECTING & OBTAINING THE DATA

To collect each dog's data, we had to enter the URL of each dog profile and scrape data.

Following multiple requests to the same server we encountered recurring glitches and finally managed to get all the required data.

	breed name	Dog Breed Group	Height	Weight	Life Span	Adaptability	Well To Apartment Living	For Novice Owners	Sensitivity Level	Tolerates Being Alone	... Intelligence	Potential For Mouthiness	Prey Drive	Tendency To Bark Or Howl
0	Afador	Mixed Breed Dogs	20 to 29 inches	50 to 75 pounds	10 to 12 years	2	1	1	3	3	5	4	4.0	4.0
1	Affenhuahua	Mixed Breed Dogs	6 to 12 inches	4 to 12 pounds	13 to 18 years	3	4	4	4	1	3	4	2.0	2.0
2	Affenpinscher	Companion Dogs	9 to 11 inches tall at the shoulder	7 to 9 pounds	12 to 14 years	3	5	4	3	1	4	4	3.0	2.0
3	Afghan Hound	Hound Dogs	24 to 26 inches tall at the shoulder	50 to 60 pounds	10 to 12 years	4	5	3	5	2	4	3	5.0	2.0
4	Airedale Terrier	Terrier Dogs	21 to 23 inches tall at the shoulder	40 to 65 pounds	10 to 13 years	2	1	2	3	2	5	5	5.0	4.0
...
...	...	Mixed Breed Dogs	12 to 20 inches	20 to 45 pounds	12 to 15 years	3	4	1	3	1	4	1	2.0	1.0

	Dog Breed	URL dog breed
0	Afador	https://dogtime.com/dog-breeds/afador
1	Affenhuahua	https://dogtime.com/dog-breeds/affenhuahua
2	Affenpinscher	https://dogtime.com/dog-breeds/affenpinscher
3	Afghan Hound	https://dogtime.com/dog-breeds/afghan-hound
4	Airedale Terrier	https://dogtime.com/dog-breeds/airedale-terrier
...
384	Whoodle	https://dogtime.com/dog-breeds/whoodle
385	Wirehaired Pointing Griffon	https://dogtime.com/dog-breeds/wirehaired-poin...
386	Xoloitzcuintli	https://dogtime.com/dog-breeds/xoloitzcuintli
387	Yorkipoo	https://dogtime.com/dog-breeds/yorkipoo
388	Yorkshire Terrier	https://dogtime.com/dog-breeds/yorkshire-terrier

389 rows x 2 columns

SCRUBBING DATA

In this step, we Cleaning, formatting and filtering the data.

In addition dealing the missing data with API.

Dealing with missing values:

We completed the missing values using the API.

Cleaning the noise:

Many strings contained unnecessary signs, we switched these columns to the average between the two numbers that we extract from the string.

Formatting:

Importing every non-numeric variable into a fitted categorical data.

```
def get_average_from_str(string):
    numbers = re.findall('[0-9]+', string)
    if len(numbers) > 1:
        return ((int(numbers[0]) + int(numbers[1])) / 2)
    elif len(numbers) == 1:
        return int(numbers[0])
    else:
        return np.nan
```

SCRUBBING DATA

For each dog we checked if had missing data we sent a request to the API with the dog's name, and got json document back and pulled the missing values into the data frame.

API connection function, to get the missing values.

```
In [13]: ## API to search missing values by breed name and data type
def get_data_from_API(dtype,breed_name):
    # API Request
    API_TOKEN = '1b7eefab0-fa10-4112-b302-88c2e5716616'
    headers = { 'x-api-key': API_TOKEN }
    queryParams = { 'q': breed_name }

    response = requests.get('https://api.thedogapi.com/v1/breeds/search',params =
        # return specific missing data
        if dtype == 'weight' or dtype == 'height':
            try:
                json_data = json.loads(response.text)[0]
            except:
                print('Not Found!')
                # not found
            try:
                return json_data[dtype]['imperial']
                # '22 - 40'
            except:
                return np.nan
        else:
            try:
                json_data = json.loads(response.text)[0]
                return json_data[dtype]
                # '10 - 12 years'
```

The Dog API - Dogs as a Service

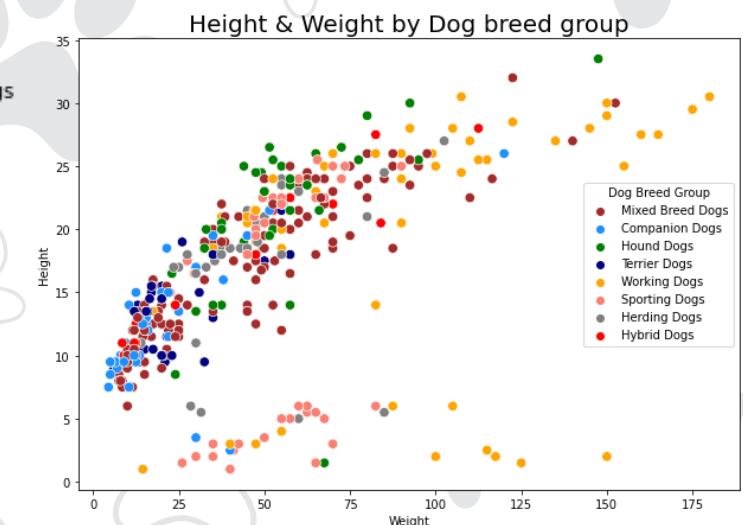
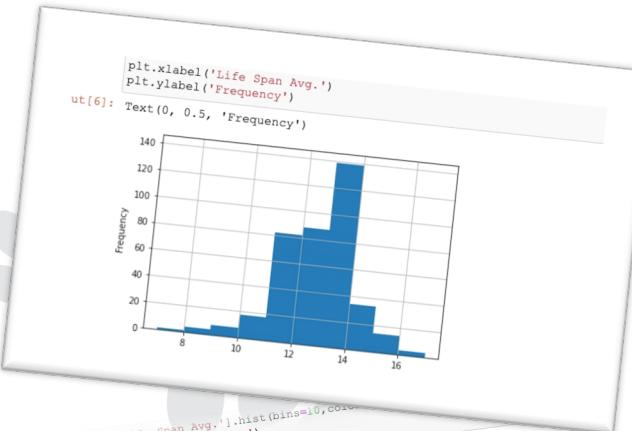
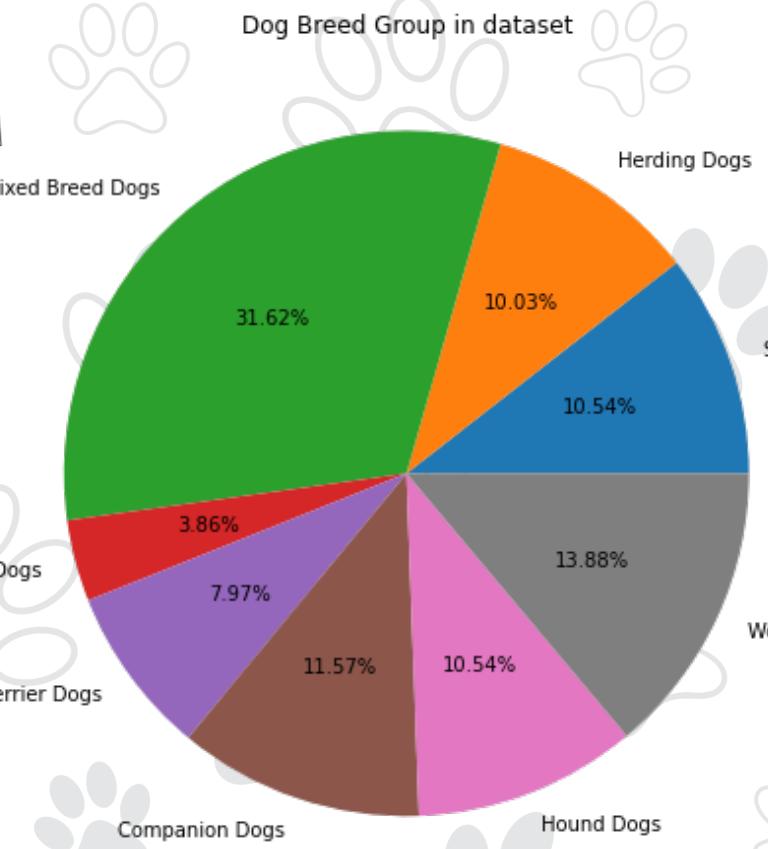
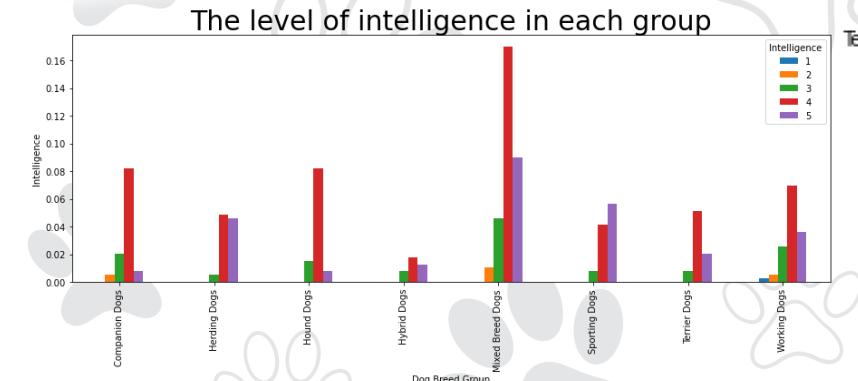
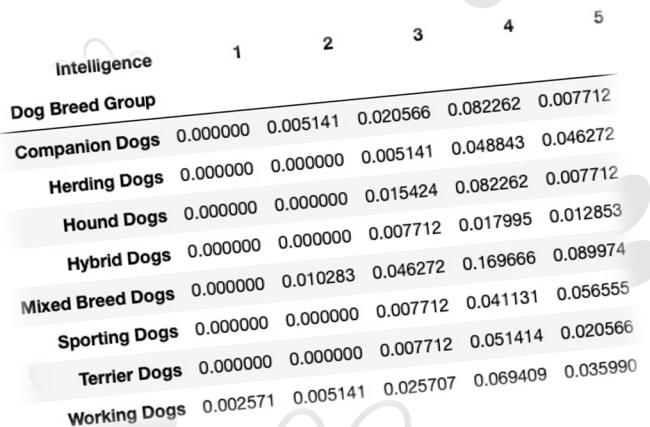
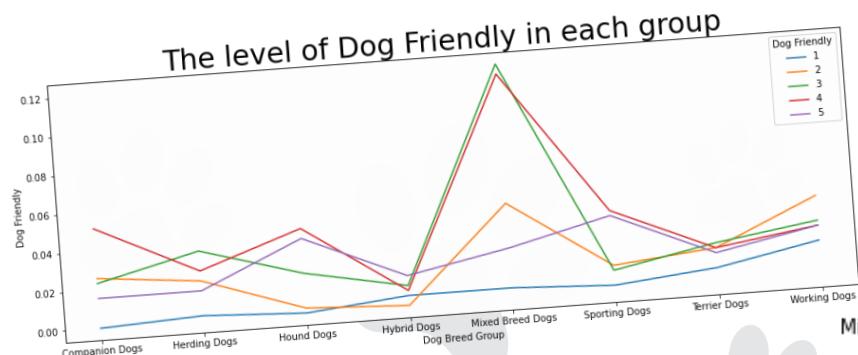
Because everyday is a Dog Day.

A public service API all about Dogs, free to use when making your fancy new App, Website or Service. [Get your license now.](#)



Source API: <https://thedogapi.com>

EXPLORING & VISUALIZING THE DATA



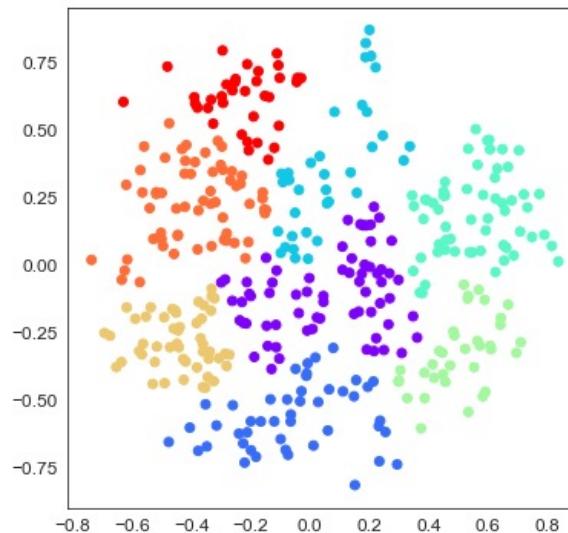
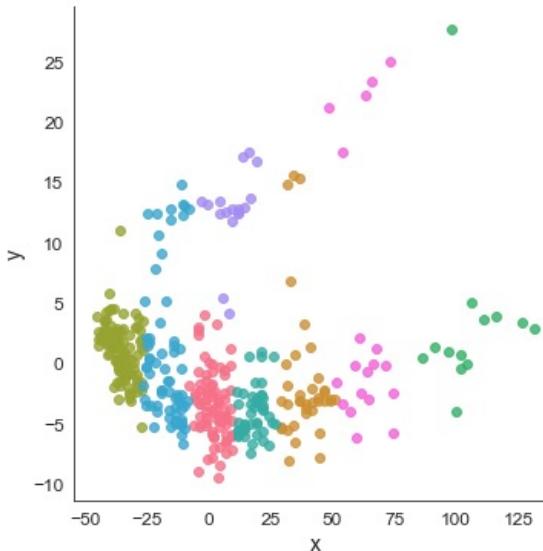
CLUSTERING AND MODELING DATA

Unsupervised Learning: Clustering by K-Means & Hierarchical clustering

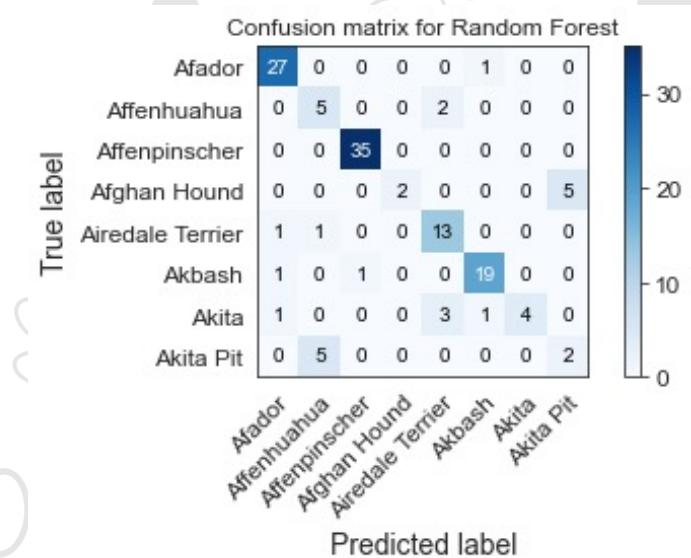
So first of all we preprocessed the data and then applied k mean clustering for different values of k and then applied agglomerative(Hierarchical clustering) clustering.

Finally we split the data into training and testing and applied random forest algorithm for classification of data and then tested on test data . Finally found accuracy.

- Visualizing the groups after using K-Means & Hierarchical clustering algorithms.
- Build a confusion matrix for the Random Forest.
- Checking the model Accuracy, how often is the classifier correct?



```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))  
Accuracy: 0.8992248062015504
```

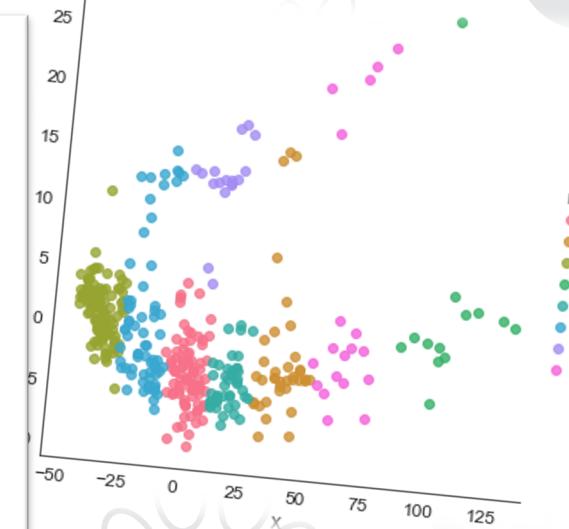


STEP 5 : INTERPRETING DATA

In this step, we compared the results of the step 4 model with the breed groups column in the original data.

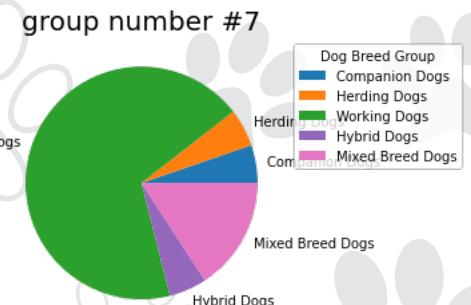
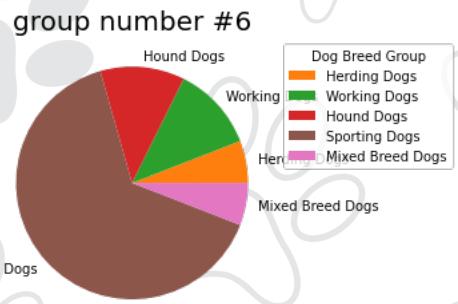
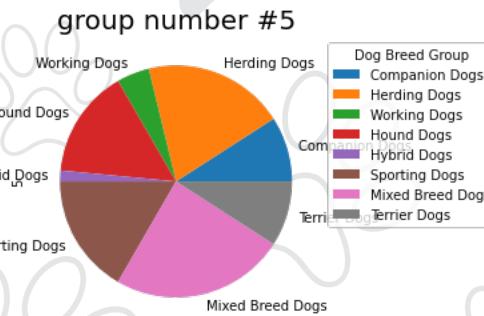
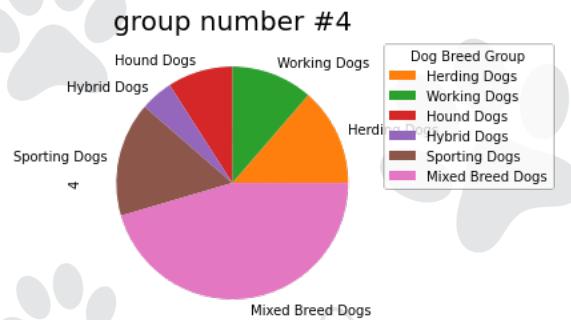
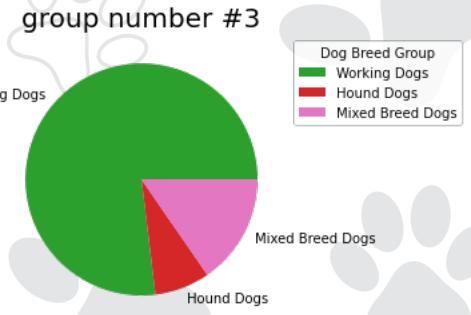
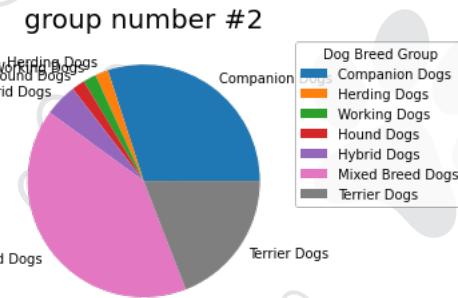
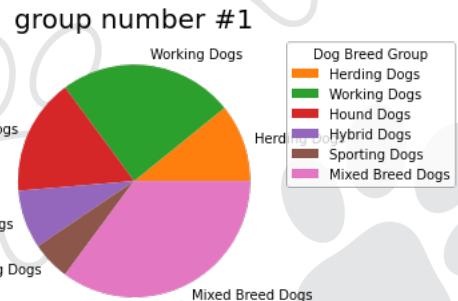
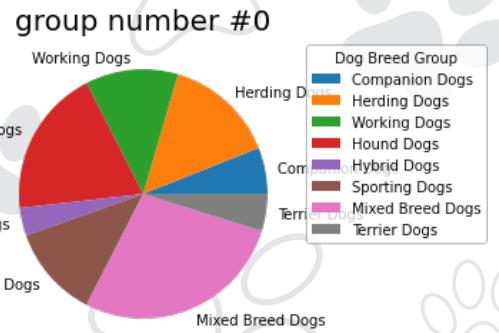
K-means algorithm split the data to 8 new groups:

	0	1	2	3	4	5	6	7
Companion Dogs	5	0	33	0	0	6	0	1
Herding Dogs	12	4	2	0	6	13	1	1
Working Dogs	10	9	2	10	5	3	2	13
Hound Dogs	16	6	2	1	4	10	2	0
Hybrid Dogs	3	3	5	0	2	1	0	1
Sporting Dogs	10	2	0	0	7	11	11	0
Mixed Breed Dogs	23	13	45	2	20	16	1	3
Terrier Dogs	4	0	21	0	0	6	0	0



STEP 5 : INTERPRETING DATA

VISUALIZING THE NEW GROUPS AFTER CLUSTERING

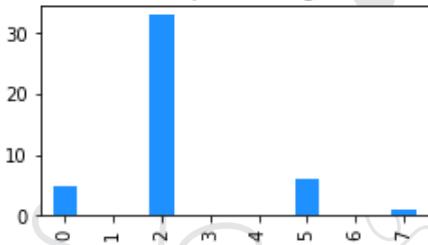


STEP 5 : INTERPRETING DATA

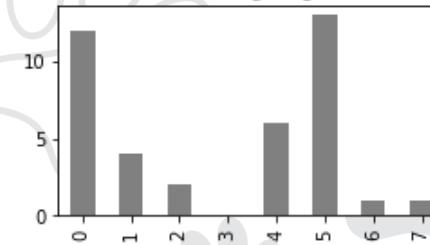
Visualizing a graph for each breed groups by the group after clustering

Graph of each breed group by clustering

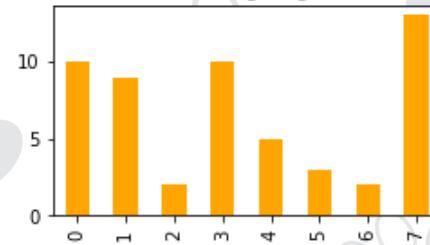
Companion Dogs



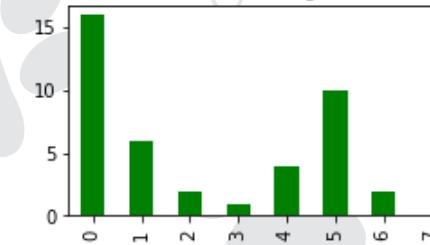
Herding Dogs



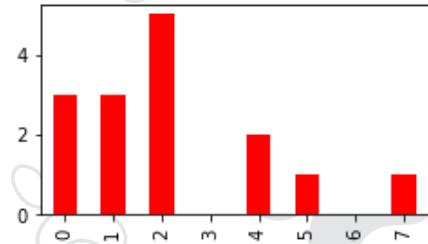
Working Dogs



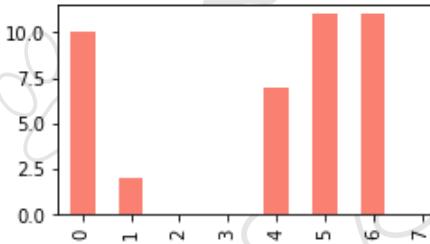
Hound Dogs



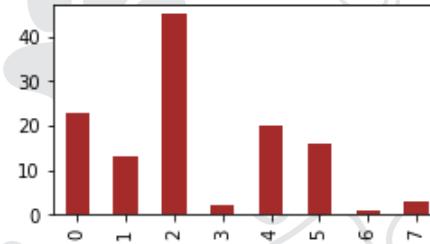
Hybrid Dogs



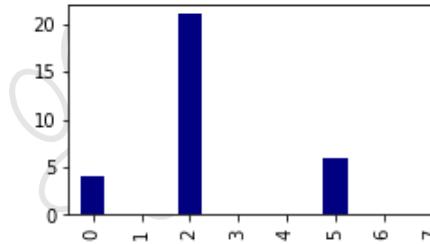
Sporting Dogs



Mixed Breed Dogs



Terrier Dogs



CONCLUSIONS

- As we can see in the slide before, the new division by the clustering largely preserves an organic grouping of dogs from the same original group.
- The similarity can be seen in the dominant features before and after clustering.

