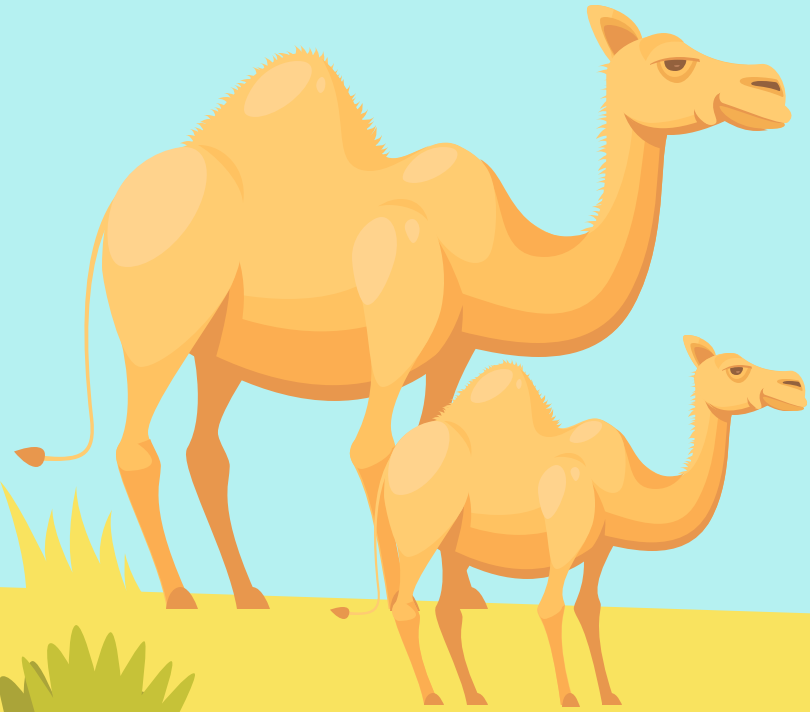
Đồ án cuối kì
Phân tích thế giới động vật

Môn học: Nhập môn khoa học dữ liệu

Nhóm 18



Giáo viên hướng dẫn

Thầy Lê Ngọc Thành
Thầy Lê Nhựt Nam
Thầy Nguyễn Thái Vũ
Thầy Trần Đại Chí
Thầy Nguyễn Bảo Long

THÀNH VIÊN

Nhóm 18

Họ và tên	MSSV
Lê Xuân Huy	20120494
Lê Xuân Hoàng	20120089
Nguyễn Thị Ánh Tuyết	20120422
Lê Nguyễn Hải Dương	20120460

Quy trình với dữ liệu

01

Thu thập dữ liệu

Chuẩn bị dữ liệu để phân tích

02

Khám phá và tiền xử lý

Trả lời một số câu hỏi để hiểu dữ liệu hơn + tiền xử lí

03

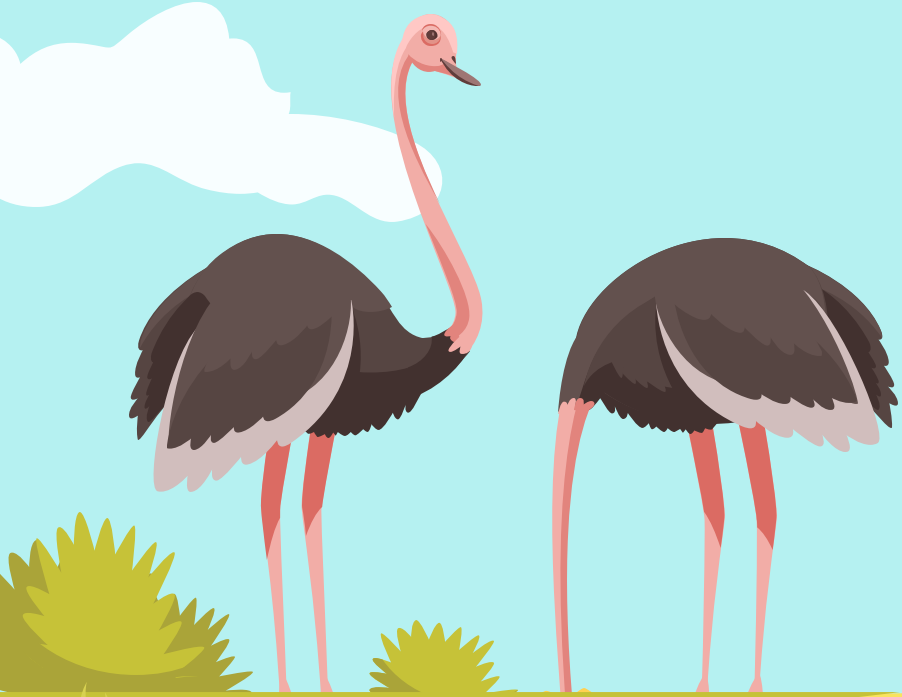
Đặt và trả lời câu hỏi

Đặt những câu hỏi có thể trả lời bằng dữ liệu

04

Đánh giá, so sánh mô hình

Xây dựng mô hình học máy và đánh giá mô hình đó



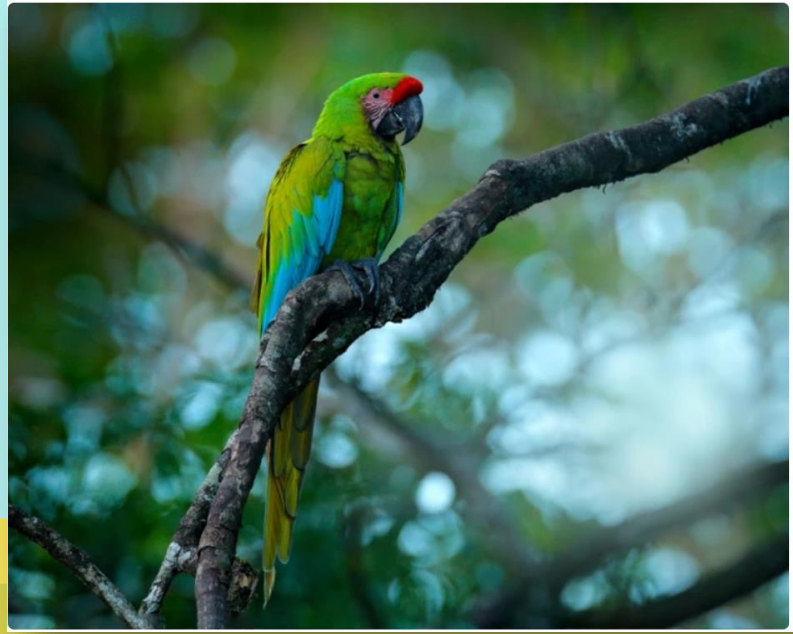
01

Thu thập dữ liệu

Chuẩn bị dữ liệu để phân tích

Thu thập dữ liệu

Dữ liệu được thu thập từ trang web: https://animalia.bio/



GREAT GREEN MACAW

Buffon's macaw, Great military macaw

KINGDOM	Animalia
PHYLUM	Chordata
CLASS	Aves
ORDER	Psittaciformes
FAMILY	Psittacidae
GENUS	Ara
SPECIES	Ara ambiguus

POPULATION SIZE	LIFE SPAN
500-1,000	60-70 YRS
WEIGHT	LENGTH
1.3 KG	85-90 CM

Đôi nét về animalia.bio

Lý do chọn chủ đề:

- Nhóm có sự hứng thú và yêu thích với thế giới động vật nên muốn tìm hiểu để biết thêm nhiều thông tin thú vị về động vật

- Biết thêm nhiều kiến thức về động vật để có thể áp dụng sau này.

Các bước thu thập dữ liệu

01

Lấy url các loài động vật

Dùng selenium, beautiful soup lấy url các loài động vật rồi lưu vào file csv

02

Chọn ra các thông tin cần thiết

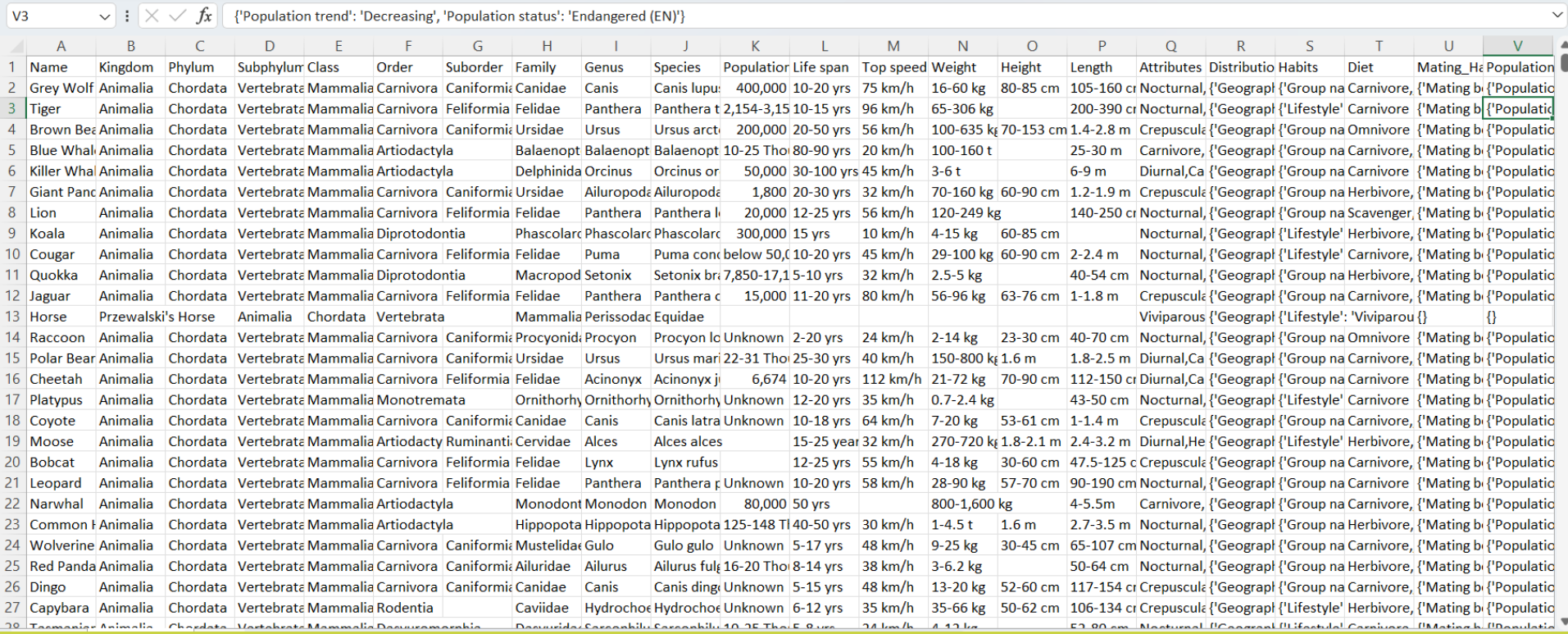
Chọn ra các thông tin có ích

03

Lấy các thông tin đã chọn

Dùng scrapy để lấy dữ liệu và lưu vào file

Dữ liệu sau khi cào về từ web



V3 {'Population trend': 'Decreasing', 'Population status': 'Endangered (EN)'}

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	Name	Kingdom	Phylum	Subphylum	Class	Order	Suborder	Family	Genus	Species	Population	Life span	Top speed	Weight	Height	Length	Attributes	Distributio	Habits	Diet	Mating_Ha	Population
2	Grey Wolf	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Canidae	Canis	Canis lupu	400,000	10-20 yrs	75 km/h	16-60 kg	80-85 cm	105-160 cr	Nocturnal,	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
3	Tiger	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Panthera	Panthera t	2,154-3,15	10-15 yrs	96 km/h	65-306 kg		200-390 cr	Nocturnal,	{'Geograpl	{'Lifestyle'	Carnivore	{'Mating b	{'Populatio
4	Brown Bea	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Ursidae	Ursus	Ursus arct	200,000	20-50 yrs	56 km/h	100-635 kg	70-153 cm	1.4-2.8 m	Crepuscula	{'Geograpl	{'Group na	Omnivore	{'Mating b	{'Populatio
5	Blue Whal	Animalia	Chordata	Vertebrata	Mammalia	Artiodactyla		Balaenopt	Balaenopt	Balaenopt	10-25 Tho	80-90 yrs	20 km/h	100-160 t		25-30 m	Carnivore,	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
6	Killer Whal	Animalia	Chordata	Vertebrata	Mammalia	Artiodactyla		Delphinida	Orcinus	Orcinus or	50,000	30-100 yrs	45 km/h	3-6 t		6-9 m	Diurnal,Ca	{'Geograpl	{'Group na	Carnivore	{'Mating b	{'Populatio
7	Giant Panc	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Ursidae	Ailuropod	Ailuropod	1,800	20-30 yrs	32 km/h	70-160 kg	60-90 cm	1.2-1.9 m	Crepuscula	{'Geograpl	{'Group na	Herbivore,	{'Mating b	{'Populatio
8	Lion	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Panthera	Panthera l	20,000	12-25 yrs	56 km/h	120-249 kg		140-250 cr	Nocturnal,	{'Geograpl	{'Group na	Scavenger,	{'Mating b	{'Populatio
9	Koala	Animalia	Chordata	Vertebrata	Mammalia	Diprotodontia		Phascolarc	Phascolarc	Phascolarc	300,000	15 yrs	10 km/h	4-15 kg	60-85 cm		Nocturnal,	{'Geograpl	{'Lifestyle'	Herbivore,	{'Mating b	{'Populatio
10	Cougar	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Puma	Puma con	below 50,(	10-20 yrs	45 km/h	29-100 kg	60-90 cm	2-2.4 m	Nocturnal,	{'Geograpl	{'Lifestyle'	Carnivore,	{'Mating b	{'Populatio
11	Quokka	Animalia	Chordata	Vertebrata	Mammalia	Diprotodontia		Macropod	Setonix	Setonix br	7,850-17,1	5-10 yrs	32 km/h	2.5-5 kg		40-54 cm	Nocturnal,	{'Geograpl	{'Group na	Herbivore,	{'Mating b	{'Populatio
12	Jaguar	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Panthera	Panthera c	15,000	11-20 yrs	80 km/h	56-96 kg	63-76 cm	1-1.8 m	Crepuscula	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
13	Horse	Przewalski's Horse	Animalia	Chordata	Vertebrata			Mammalia	Perissodac	Equidae							Viviparous	{'Geograpl	{'Lifestyle': 'Viviparou	{}		{}
14	Raccoon	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Procyonid	Procyon	Procyon lo	Unknown	2-20 yrs	24 km/h	2-14 kg	23-30 cm	40-70 cm	Nocturnal,	{'Geograpl	{'Group na	Omnivore	{'Mating b	{'Populatio
15	Polar Bear	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Ursidae	Ursus	Ursus mari	22-31 Tho	25-30 yrs	40 km/h	150-800 kg	1.6 m	1.8-2.5 m	Diurnal,Ca	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
16	Cheetah	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Acinonyx	Acinonyx j	6,674	10-20 yrs	112 km/h	21-72 kg	70-90 cm	112-150 cr	Diurnal,Ca	{'Geograpl	{'Group na	Carnivore	{'Mating b	{'Populatio
17	Platypus	Animalia	Chordata	Vertebrata	Mammalia	Monotremata		Ornithorhy	Ornithorhy	Ornithorhy	Unknown	12-20 yrs	35 km/h	0.7-2.4 kg		43-50 cm	Nocturnal,	{'Geograpl	{'Lifestyle'	Carnivore	{'Mating b	{'Populatio
18	Coyote	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Canidae	Canis	Canis latra	Unknown	10-18 yrs	64 km/h	7-20 kg	53-61 cm	1-1.4 m	Crepuscula	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
19	Moose	Animalia	Chordata	Vertebrata	Mammalia	Artiodacty	Ruminanti	Cervidae	Alces	Alces alces		15-25 year	32 km/h	270-720 kg	1.8-2.1 m	2.4-3.2 m	Diurnal,He	{'Geograpl	{'Lifestyle'	Herbivore,	{'Mating b	{'Populatio
20	Bobcat	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Lynx	Lynx rufus		12-25 yrs	55 km/h	4-18 kg	30-60 cm	47.5-125 c	Crepuscula	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
21	Leopard	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Feliformia	Felidae	Panthera	Panthera p	Unknown	10-20 yrs	58 km/h	28-90 kg	57-70 cm	90-190 cm	Nocturnal,	{'Geograpl	{'Group na	Carnivore	{'Mating b	{'Populatio
22	Narwhal	Animalia	Chordata	Vertebrata	Mammalia	Artiodactyla		Monodont	Monodon	Monodon	80,000	50 yrs		800-1,600 kg		4-5.5m	Carnivore,	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
23	Common H	Animalia	Chordata	Vertebrata	Mammalia	Artiodactyla		Hippopota	Hippopota	Hippopota	125-148 Tl	40-50 yrs	30 km/h	1-4.5 t	1.6 m	2.7-3.5 m	Nocturnal,	{'Geograpl	{'Group na	Herbivore,	{'Mating b	{'Populatio
24	Wolverine	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Mustelida	Gulo	Gulo gulo	Unknown	5-17 yrs	48 km/h	9-25 kg	30-45 cm	65-107 cm	Nocturnal,	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
25	Red Panda	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Ailuridae	Ailurus	Ailurus ful	16-20 Tho	8-14 yrs	38 km/h	3-6.2 kg		50-64 cm	Nocturnal,	{'Geograpl	{'Group na	Herbivore,	{'Mating b	{'Populatio
26	Dingo	Animalia	Chordata	Vertebrata	Mammalia	Carnivora	Californi	Canidae	Canis	Canis ding	Unknown	5-15 yrs	48 km/h	13-20 kg	52-60 cm	117-154 cr	Crepuscula	{'Geograpl	{'Group na	Carnivore,	{'Mating b	{'Populatio
27	Capybara	Animalia	Chordata	Vertebrata	Mammalia	Rodentia		Caviidae	Hydrochoe	Hydrochoe	Unknown	6-12 yrs	35 km/h	35-66 kg	50-62 cm	106-134 cr	Crepuscula	{'Geograpl	{'Lifestyle'	Herbivore,	{'Mating b	{'Populatio
28	Tasmanian	Animalia	Chordata	Vertebrata	Mammalia	Dasyuromorphia		Dasyuridae	Sarcophilu	Sarcophilu	10-25 Tho	5-8 yrs	24 km/h	4-12 kg		52-80 cm	Nocturnal,	{'Geograpl	{'Lifestyle'	Carnivore,	{'Mating b	{'Populatio

02

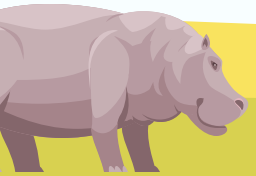
Khám phá và tiền xử lý dữ liệu

Trả lời 1 số câu hỏi cần thiết để hiểu dữ liệu hơn,
Xử lý 1 số cột có kiểu dữ liệu chưa phù hợp

Khám phá dữ liệu

Để hiểu hơn về dữ liệu, có một số câu hỏi cần phải trả lời như sau:

- Dữ liệu có bao nhiêu dòng và bao nhiêu cột?
- Mỗi dòng có ý nghĩa gì? Có vấn đề các dòng có ý nghĩa khác nhau không?
- Dữ liệu có các dòng bị lặp không?
- Mỗi cột có ý nghĩa gì?
- Mỗi cột hiện đang có kiểu dữ liệu gì?
- Các giá trị của mỗi cột được phân bố như thế nào?
- Số-lượng/tỉ-lệ các giá trị thiếu?
- Số lượng các giá trị khác nhau? Show một vài giá trị .
- Có cột nào có kiểu dữ liệu chưa phù hợp để có thể xử lý tiếp không?



Khám phá dữ liệu

Dữ liệu cũng tương đối nhiều (cả dòng và cột). Cũng khá ổn, không có dòng nào bị trùng.

Có rất nhiều cột dữ liệu kiểu phân loại.

```
animal_df.shape
```

```
(28525, 22)
```

	Attributes	Distribution	Habits	Diet	Mating_Habits	Population
	Nocturnal,Carnivore,Scavenger,Terrestrial,Curs...	{'Geography': {'Continents': 'Asia, Europe, No...	{'Group name': 'pack, route, rout', 'Lifestyle...	Carnivore,,Scavenger	{'Mating behavior': 'Monogamy', 'Reproduction ...	{'Population trend': 'Stable', 'Population sta...
	Nocturnal,Carnivore,Cursorial,Terrestrial,Ambu...	{'Geography': {'Continents': 'Asia', 'Subconti...	{'Lifestyle': 'Cursorial, Terrestrial, Ambush ...	Carnivore	{'Mating behavior': 'Polygyny', 'Reproduction ...	{'Population trend': 'Decreasing', 'Population...

Khám phá dữ liệu

Ý nghĩa của các cột trong bộ dữ liệu đã thu thập (do nhóm tự tổng hợp). Có tổng cộng 22 cột.

	Column	Description
0	Name	Name of Animal
1	Kingdom	Giới là đơn vị phân loại lớn nhất bao gồm các ngành sinh vật có chung những đặc điểm nhất định.
2	Phylum	Ngành: Một đơn vị phân loại ở cấp dưới giới và trên lớp.
3	Subphylum	Phân Ngành: là một bậc phân loại nằm trung gian giữa ngành và lớp hoặc phân thứ ngành hoặc liên lớp
4	Class	Lớp: Một mức độ phân loại động vật trong một ngành. Các lớp sau đó được chia nhỏ thành các nhóm khác được gọi là bộ
5	Order	Bộ: là một cấp nằm giữa lớp và họ.
6	Suborder	Phân bộ: một nhóm thực vật hoặc động vật có liên quan, có quan hệ gần gũi hơn so với một bộ nhưng ít giống nhau hơn so với một họ.
7	Family	Họ: Bộ được chia thành các họ, họ sẽ được chia thành các chi
8	Genus	Chi: còn gọi là giống là một đơn vị phân loại sinh học dùng để chỉ một hoặc một nhóm loài có kiểu hình tương tự và mối quan hệ tiến hóa gần gũi với nhau.
9	Species	Loài: là một nhóm các cá thể sinh vật có những đặc điểm sinh học tương đối giống nhau và có khả năng giao phối với nhau và sinh sản ra thế hệ tương lai.
10	Population size	Ước tính số cá thể đang sống của loài
11	Life span	Tuổi thọ
12	Top speed	Tốc độ
13	Weight	Cân nặng
14	Height	Chiều cao
15	Length	Chiều dài
16	Attributes	Một số thông tin chính về loài vật
17	Distribution	Phân bố: Nơi sinh sống
18	Habits	Thói quen
19	Diet	Chế độ ăn (ăn cỏ, ăn tạp,...)
20	Mating_Habits	Thói quen giao phối
21	Population	Bao gồm xu hướng dân số (tăng, giảm, ổn định) và trạng thái dân số (nguy cấp, bị đe dọa, ít quan tâm,...)

Khám phá dữ liệu

Trừ thông tin liên quan đến tên loài và họ hàng (các thông tin trong cấp bậc phân loại), đa số các thông tin tỉ lệ miss khá cao.

ubphylum	Class	Order	Suborder	Family	Genus	Species	...	Top speed	Weight	Height	Length	Attributes	Distribution	Habits	Diet
64.4	0.0	1.5	91.0	0.0	0.0	0.0	...	98.0	90.0	98.9	89.7	0.0	0.0	0.0	90.9

Ngoài các cột có tỉ lệ miss cao, các cột kiểu phân loại sau khi khai phá cũng có rất nhiều giá trị rỗng.

Contine	Subcon	Countri	Regions	WWT Bi	Group	Diet	Biogeog	Lifestyl	Seasona	BC_min	BC_ma
['Asia']	[]	['Myanma	[]	[]	[]	[]	['Indomala	[]	[]	1	
['South Am	[]	['Ecuador'	[]	[]	[]	[]	['Neotropi	[]	[]		
['']	[]	[]	[]	[]	[]	[]	[]	['Terrestria	[]		
['']	[]	[]	[]	[]	[]	[]	[]	['Terrestria	[]		
[]	[]	['']	[]	[]	[]	[]	[]	[]	[]		
[]	[]	['']	[]	[]	[]	[]	[]	[]	[]		
['South Am	[]	['Argentina	[]	[]	[]	[]	['Neotropi	[]	[]		
[]	[]	[]	[]	[]	[]	[]	['Indomala	[]	[]	1	
[]	[]	['']	[]	[]	[]	[]	[]	[]	[]		

Tiền xử lí dữ liệu

Life span	Top speed	Weight	Height	Length
10-20 yrs	75 km/h	16-60 kg	80-85 cm	105-160 cm
10-15 yrs	96 km/h	65-306 kg	NaN	200-390 cm
20-50 yrs	56 km/h	100-635 kg	70-153 cm	1.4-2.8 m
80-90 yrs	20 km/h	100-160 t	NaN	25-30 m

Có một số cột kiểu dữ liệu chưa phù hợp như “Life span”, “Top speed”, “Weight”, “Height”, “Length” có nhiều loại hình thức và nhiều loại đơn vị.

Tiền xử lí dữ liệu

Life span	Top speed	Weight	Height	Length
15.0	75.0	38.0	0.825	1.325
12.5	96.0	185.5	NaN	2.950
35.0	56.0	367.5	1.115	2.100
85.0	20.0	130000.0	NaN	27.500

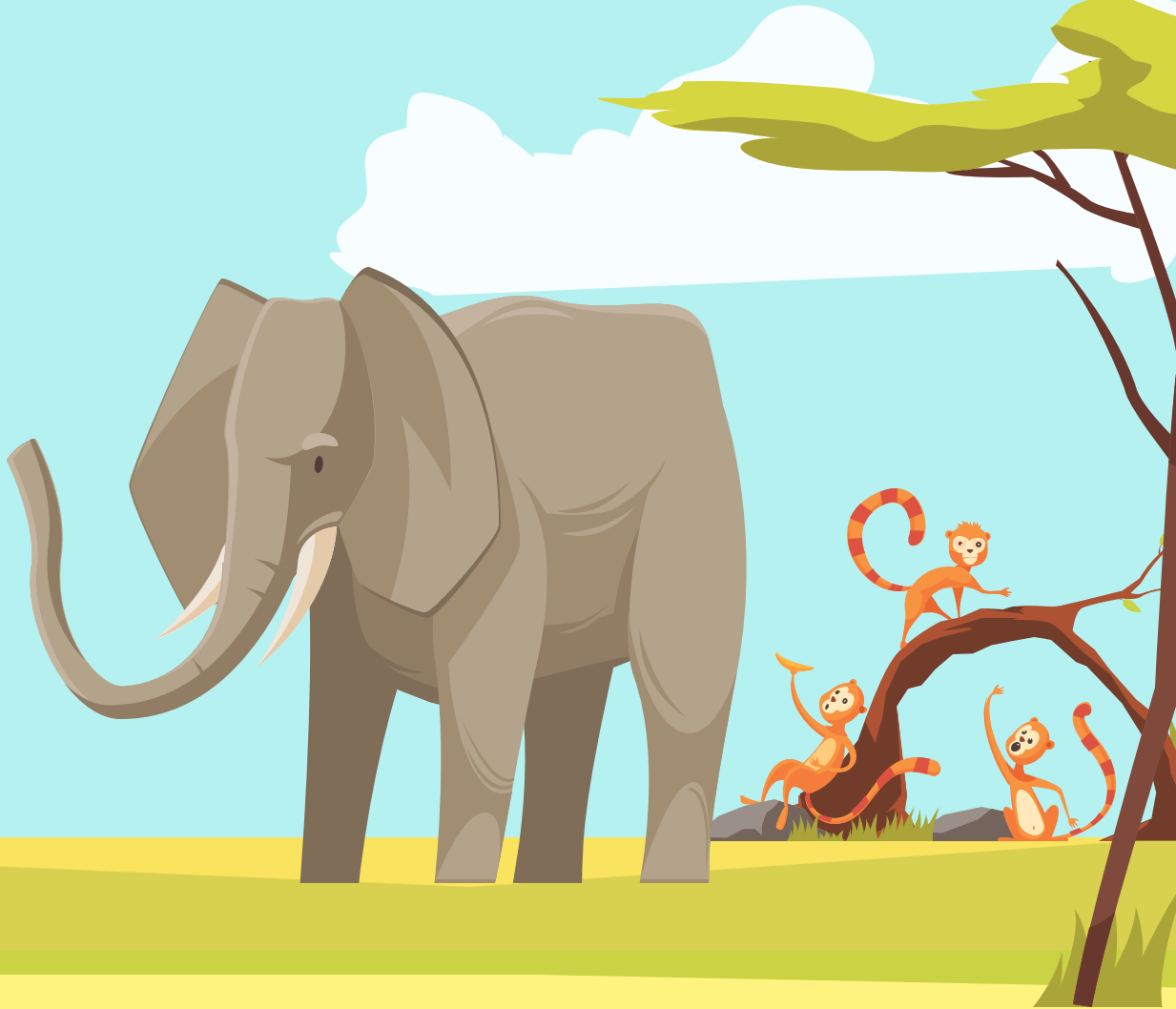
Sau khi xử lí, giá trị các cột đó sẽ như sau:

Quy về cùng 1 đơn vị:
- m (mét) cho “Height”, “Length”
- kg cho “Weight”
- km/h cho “Top speed”

03

Đặt và trả lời câu hỏi

Đặt và trả lời những câu hỏi có thể trả lời bằng dữ liệu



Đưa ra các câu hỏi

01

Xu hướng phát triển các loài

02

Sự thay đổi của Attributes và nơi ở như thế nào?

03

Động vật thích sống ở đâu nhất?

04

Chủng loài nào xuất hiện nhiều nơi trên trái đất nhất

05

Lớp, Bộ, Họ, Chi, Loài và ý nghĩa. Số lượng nhóm cá thể ứng với các bậc.

06

Quan hệ giữa chiều dài, cân nặng, chiều cao với tốc độ

1. Xu hướng phát triển các loài

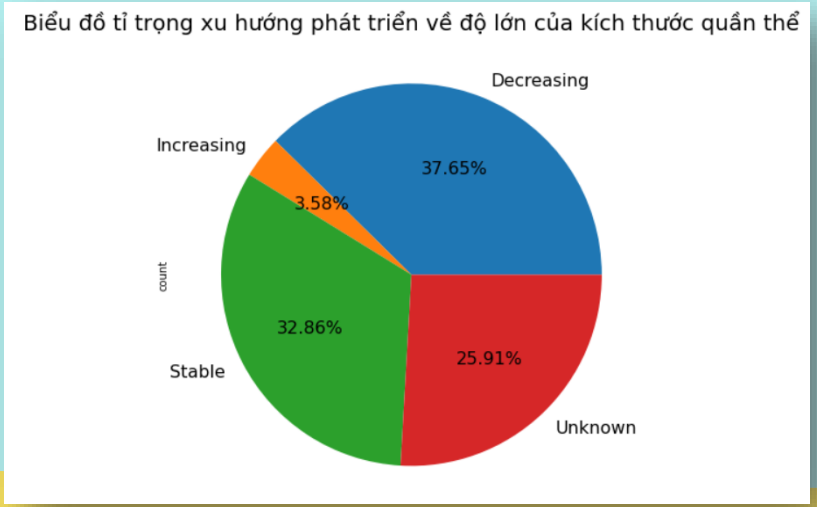
Trả lời câu hỏi này giúp chúng ta biết được xu hướng phát triển và nguy cơ tuyệt chủng của các loài.

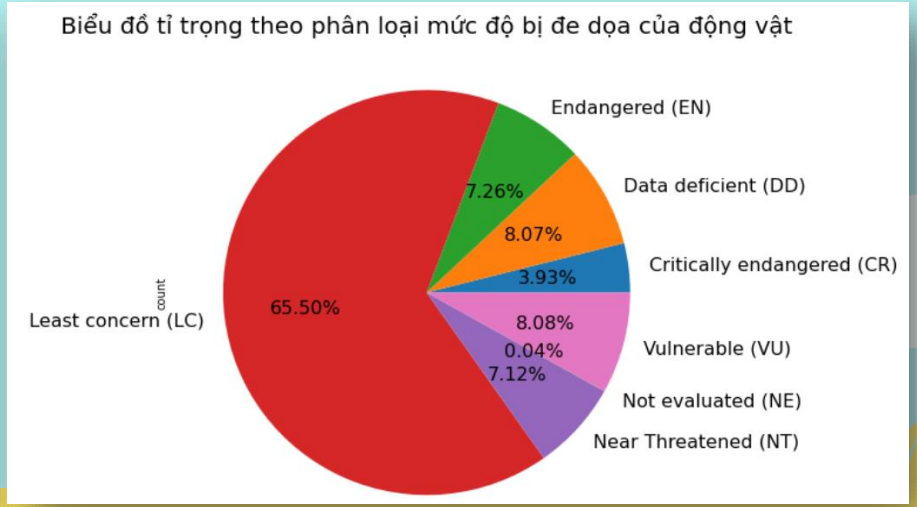
Để trả lời câu hỏi này, chúng ta sẽ cần làm rõ 1 số vấn đề sau:

- Số lượng loài theo xu hướng phát triển về kích thước quần thể (tăng, giảm, hay ổn định?)
- Số lượng loài theo phân loại (Least concern (LC), Extinct (EX), Critically endangered (CR),...)
- Những loài vật đang giảm có phải đang bị đe dọa không? Những loài bị đe dọa có phải đều đang giảm không?

1. Xu hướng phát triển các loài

Để làm rõ các vấn đề đã nêu ở trên, chúng ta tiến hành vẽ 1 số biểu đồ, sau đó là có thể rút ra nhận xét để kết luận.

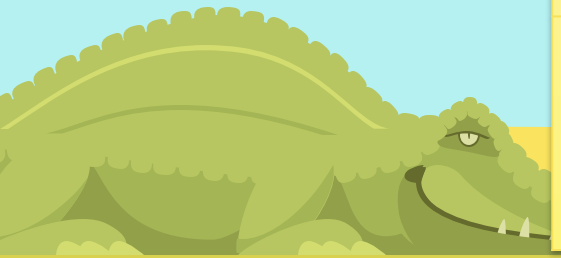




1. Xu hướng phát triển các loài

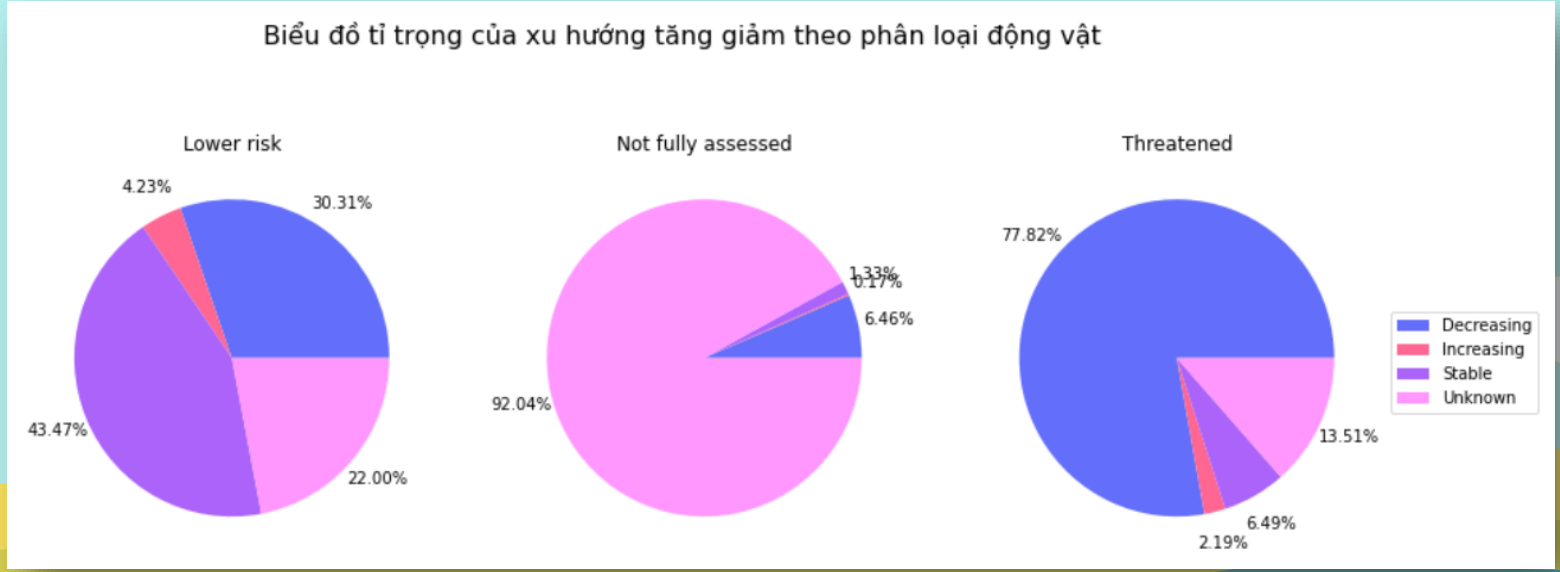
Để dễ dàng phân loại hơn, ta đưa các phân loại mức độ nguy hiểm thành các loại mới như sau. Chúng ta tự làm tay, lưu vào file csv rồi khi cần xử lý thì sẽ đọc lên dataframe.

new_status	status
Lower risk	Least concern (LC)
	Near Threatened (NT)
Not fully assessed	Data deficient (DD)
	Not Evaluated (NE)
Threatened	Endangered (EN)
	Critically endangered (CR)
	Vulnerable (VU)



1. Xu hướng phát triển các loài

Để làm rõ các vấn đề đã nêu ở trên, chúng ta tiến hành vẽ 1 số biểu đồ, sau đó là có thể rút ra nhận xét để kết luận.

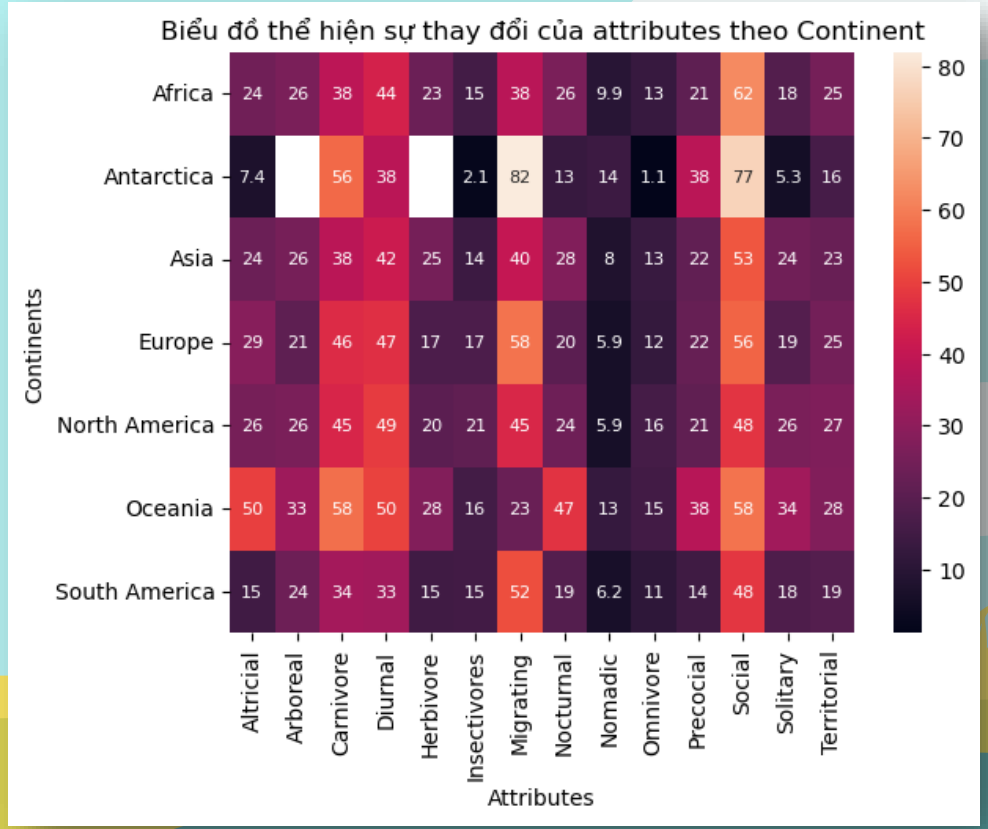


2. Sự thay đổi của Attributes và nơi ở như thế nào?

Theo continent

Màu sắc trong hình biểu thị cho phần trăm số động vật mang thuộc tính đó (màu trắng là 0%)

Hình bên cho thấy được sự khác nhau của 1 thuộc tính giữa các châu lục và các thuộc tính khác nhau trong cùng 1 châu lục

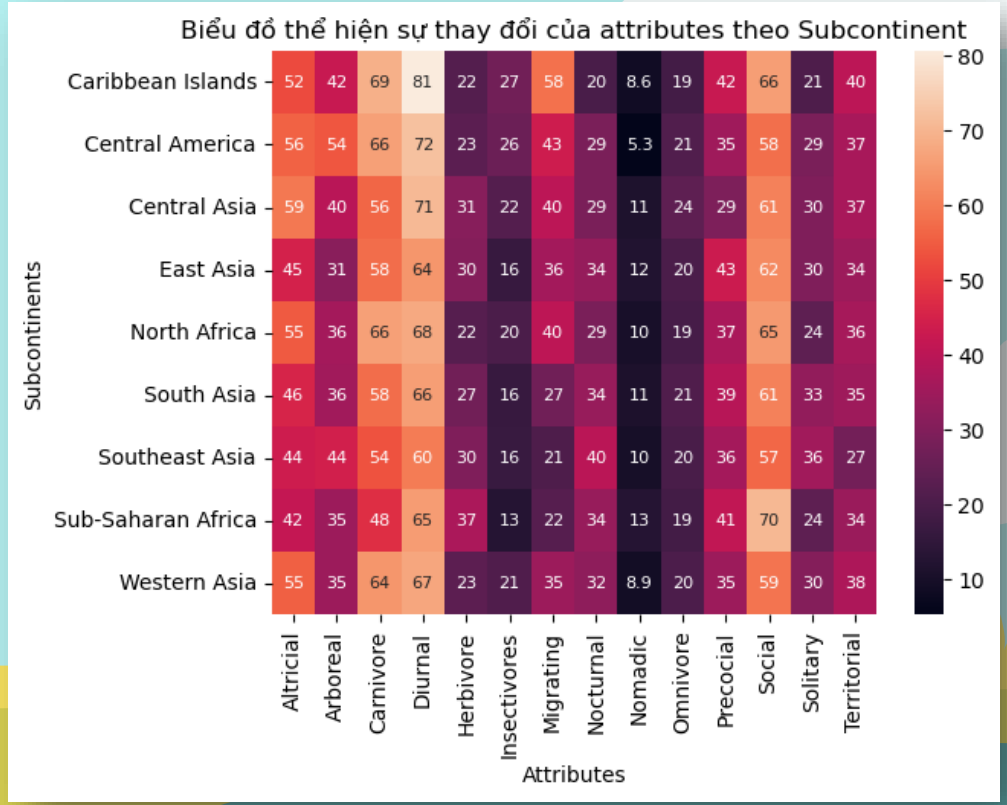


2. Sự thay đổi của Attributes và nơi ở như thế nào?

Theo Subcontinent

Màu sắc trong hình biểu thị cho phần trăm số động vật mang thuộc tính đó (màu trắng là 0%)

Hình bên cho thấy được sự khác nhau của 1 thuộc tính giữa vùng miền

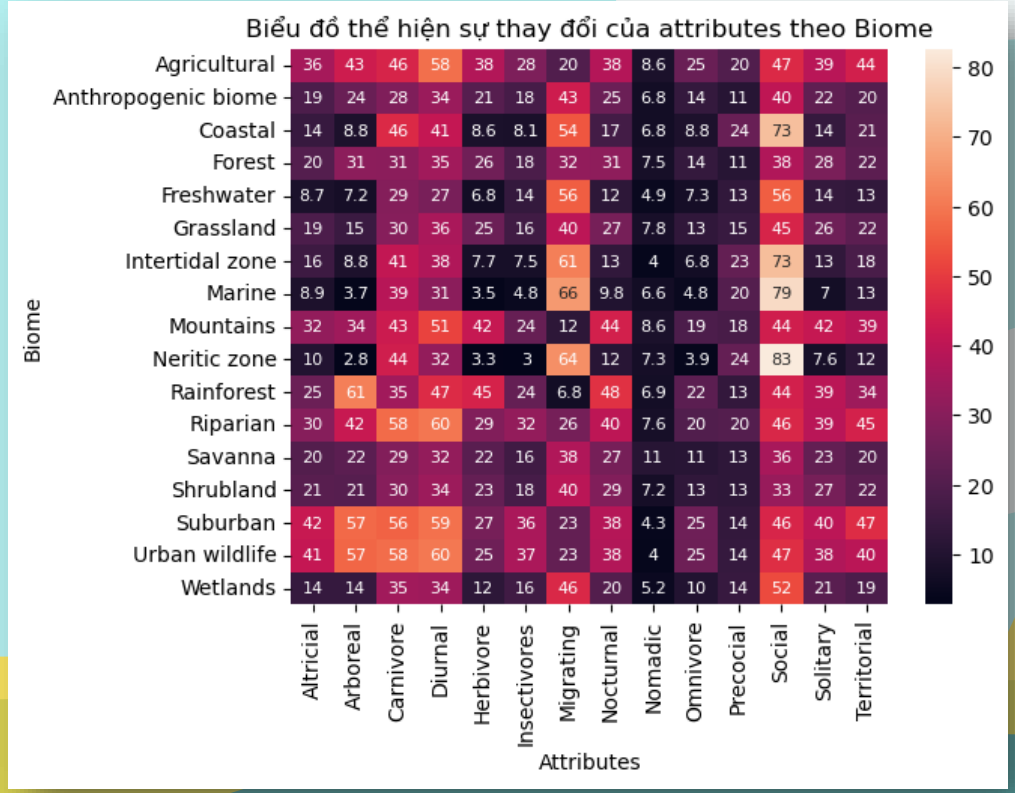


2. Sự thay đổi của Attributes và nơi ở như thế nào?

Theo biome

Màu sắc trong hình biểu thị cho phần trăm số động vật mang thuộc tính đó (màu trắng là 0%)

Hình bên cho thấy được sự khác nhau của 1 thuộc tính giữa các hệ sinh thái



3. Động vật thích sống ở đâu nhất?

Trả lời câu hỏi này giúp chúng ta hiểu rõ hơn về môi trường, khí hậu ưa thích của các loài động vật, những nơi như thế nào thì có động vật phong phú, nơi nào thì là điều kiện sống khắc nghiệt đối với động vật.

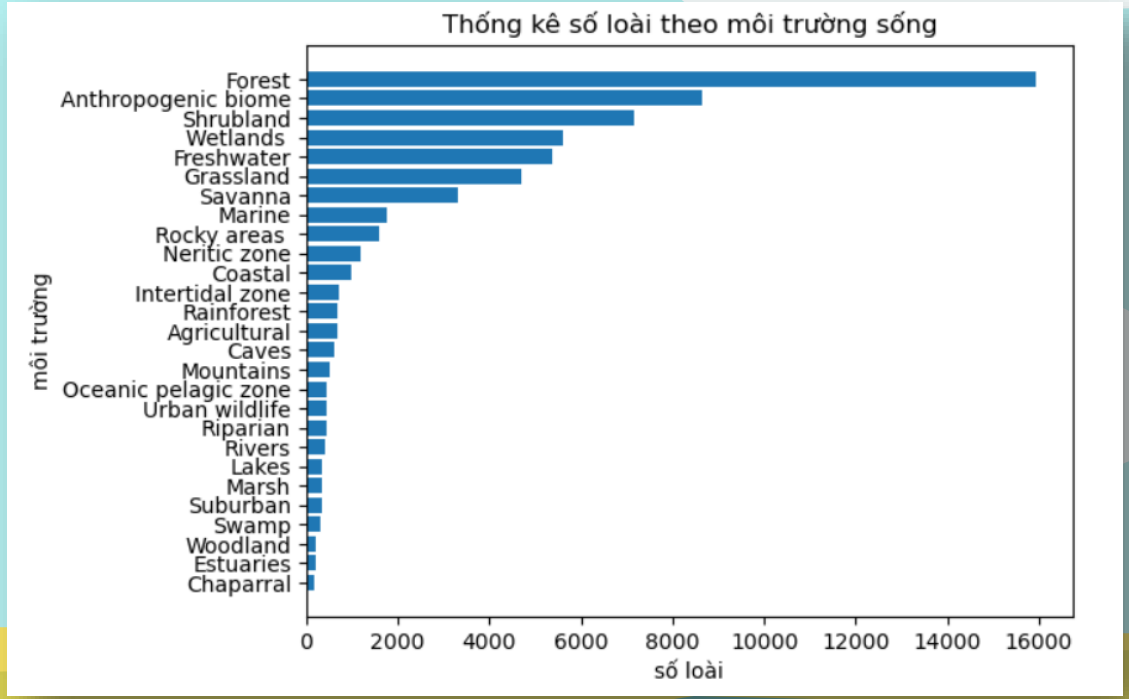
Dựa vào các thông tin mà bộ dữ liệu cung cấp, chúng ta có thể trả lời câu hỏi theo hướng: môi trường sống và khí hậu. (Thông tin ở 2 cột “Biome “ và “Climate”)

Thông tin ở 2 cột “Biome” và “Climate” đều ở dạng list, thế nên trước đó chúng ta cần có một vài thao tác nhỏ để xử lý, explode các giá trị trong cột ra, sau đó tiến hành trực quan hóa kết quả.

3. Động vật thích sống ở đâu nhất?

Qua biểu đồ ta có nhận xét:

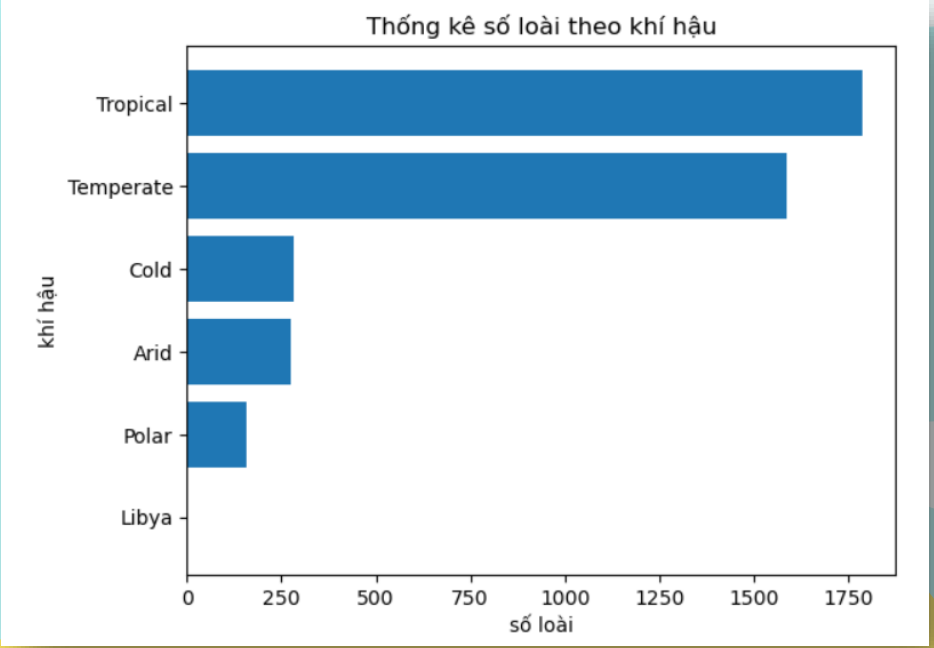
- Môi trường sống có nhiều loại động vật sống nhất là trong rừng, muốn gặp được nhiều loại động vật đa dạng thì nơi tốt nhất sẽ là trong các khu rừng.
- Con người cũng đã sống chung được với rất nhiều loài động vật (Theo bộ dữ liệu này), các loài động vật khác trong tự nhiên thường sống ở trong rừng, cây bụi, đất ngập nước, trong nước ngọt, đồng cỏ, thảo nguyên.



3. Động vật thích sống ở đâu nhất?

Có vẻ cột này dữ liệu bị miss khá nhiều, nhưng số dữ liệu còn lại cũng không phải là quá ít, cũng có thể dùng.

Qua biểu đồ này, chúng ta liền có thể nhìn ra các loại động vật thích sống ở vùng khí hậu nhiệt đới và ôn đới. Mà điều này cũng không lấy làm lạ, rõ là các kiểu khí hậu quá mức lạnh giá hay khô khan rất khắc nghiệt, khó có nhiều loài động vật sinh sống.



4. Chủng loài nào xuất hiện nhiều nơi trên trái đất nhất

Trả lời câu hỏi này giúp chúng ta hiểu rõ hơn về sự phân bố của các chủng loài trên trái đất.

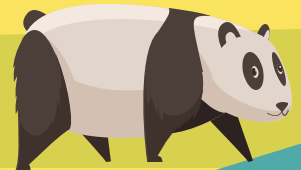
Chủng loài nói đến ở đây là bậc phân loại. Chúng em quyết định chọn bậc phân loại là Lớp và Bộ. Vậy câu hỏi của chung ta sẽ là Lớp (Bộ) nào xuất hiện nhiều nhất trên trái đất.

Chúng ta sẽ phân tích trên 2 mức: **Mức lục địa và Mức quốc gia**.

Kết quả của 2 mức này sẽ cho chúng ta biết Lớp (Bộ) nào xuất hiện nhiều nhất trên trái đất.

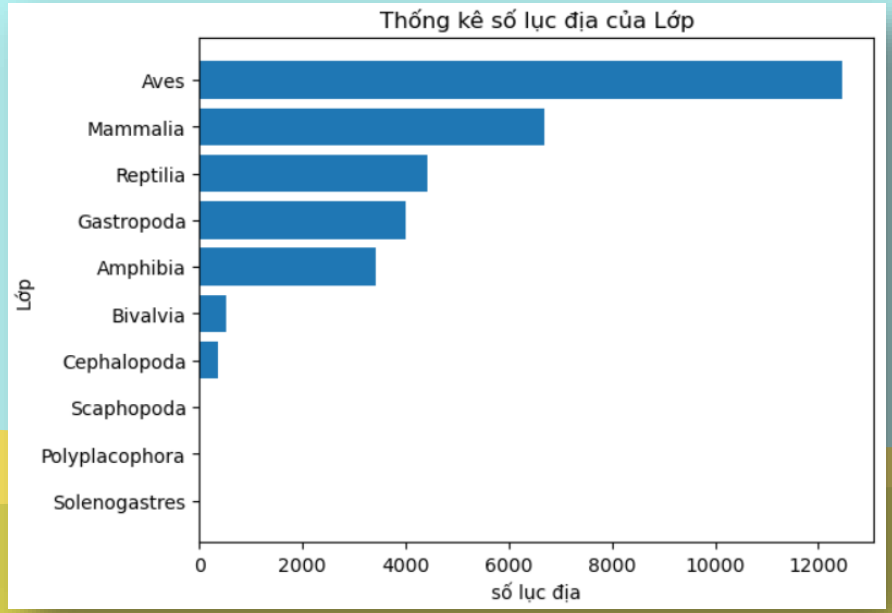
So sánh giữa 2 mức cho ta thấy được đặc điểm phân bố của Lớp (Bộ).

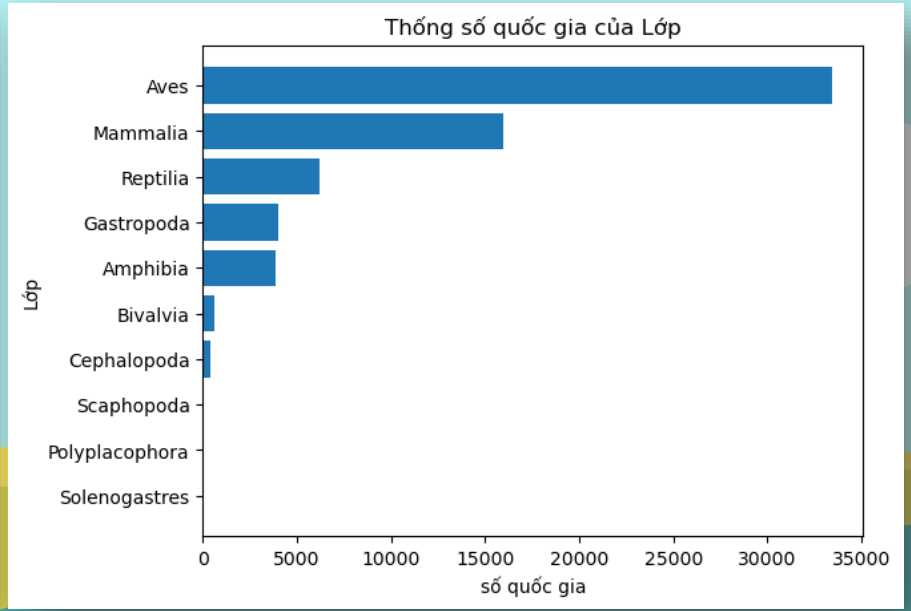
Chúng ta sẽ tiến hành thống kê tổng số nơi (lục địa hoặc quốc gia) mà Lớp (Bộ) đó xuất hiện trên trái đất để biết Lớp (Bộ) nào xuất hiện nhiều nhất trên trái đất.



4. Chủng loài nào xuất hiện nhiều nơi trên trái đất nhất

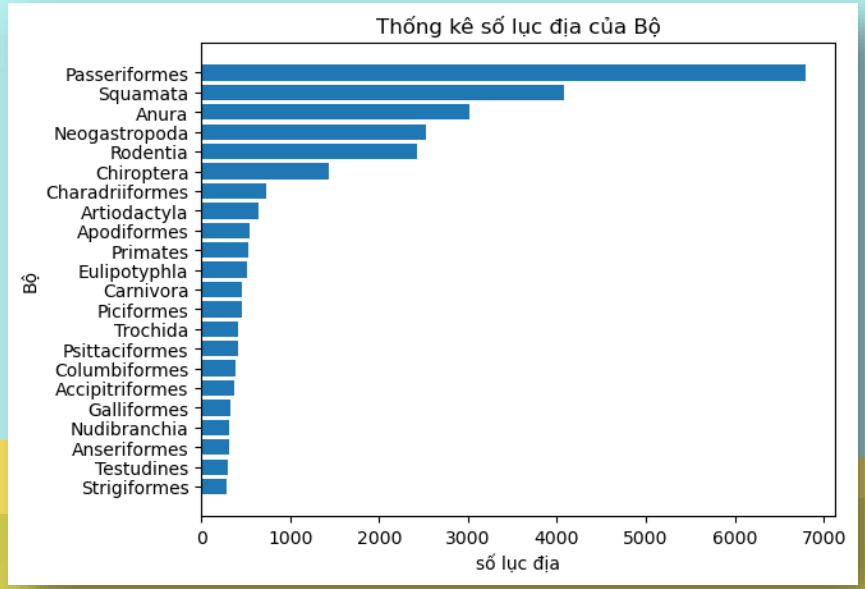
Để làm rõ các vấn đề đã nêu ở trên, chúng ta tiến hành vẽ 1 số biểu đồ, sau đó là có thể rút ra nhận xét để kết luận.

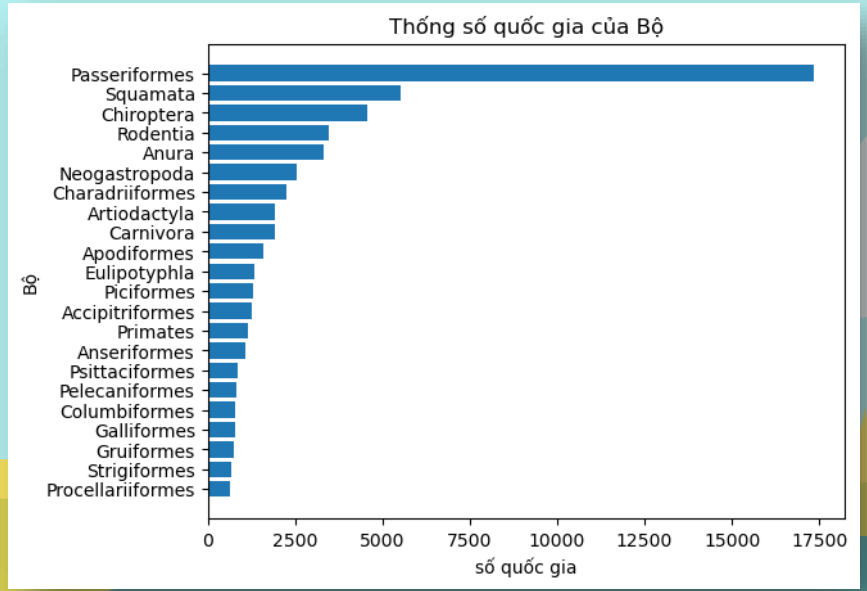




4. Chủng loài nào xuất hiện nhiều nơi trên trái đất nhất

Để làm rõ các vấn đề đã nêu ở trên, chúng ta tiến hành vẽ 1 số biểu đồ, sau đó là có thể rút ra nhận xét để kết luận.





5. Quan hệ giữa chiều dài, cân nặng, chiều cao với tốc độ

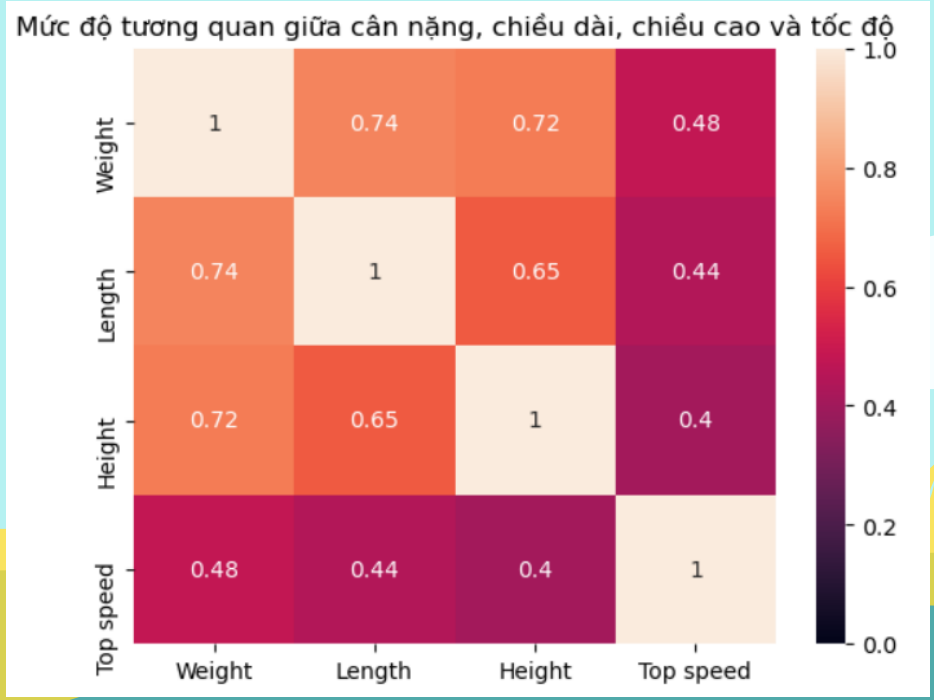
Nguồn cảm hứng của câu hỏi từ việc chúng ta hay thấy những động vật có cơ thể thon gọn, trọng lượng vừa phải (chó, mèo, hổ, báo, ngựa,...) thì thường có tốc độ nhanh hơn so với những loài động vật kích thước lớn như hà mã, voi, gấu trúc,... Vì vậy chúng ta có thể đặt ra câu hỏi là kích thước và trọng lượng có mối liên hệ với tốc độ.

5. Quan hệ giữa chiều dài, cân nặng, chiều cao với tốc độ

Đầu tiên, cần xử lí các outliers ở các cột chiều dài, cân nặng, chiều cao và tốc độ.

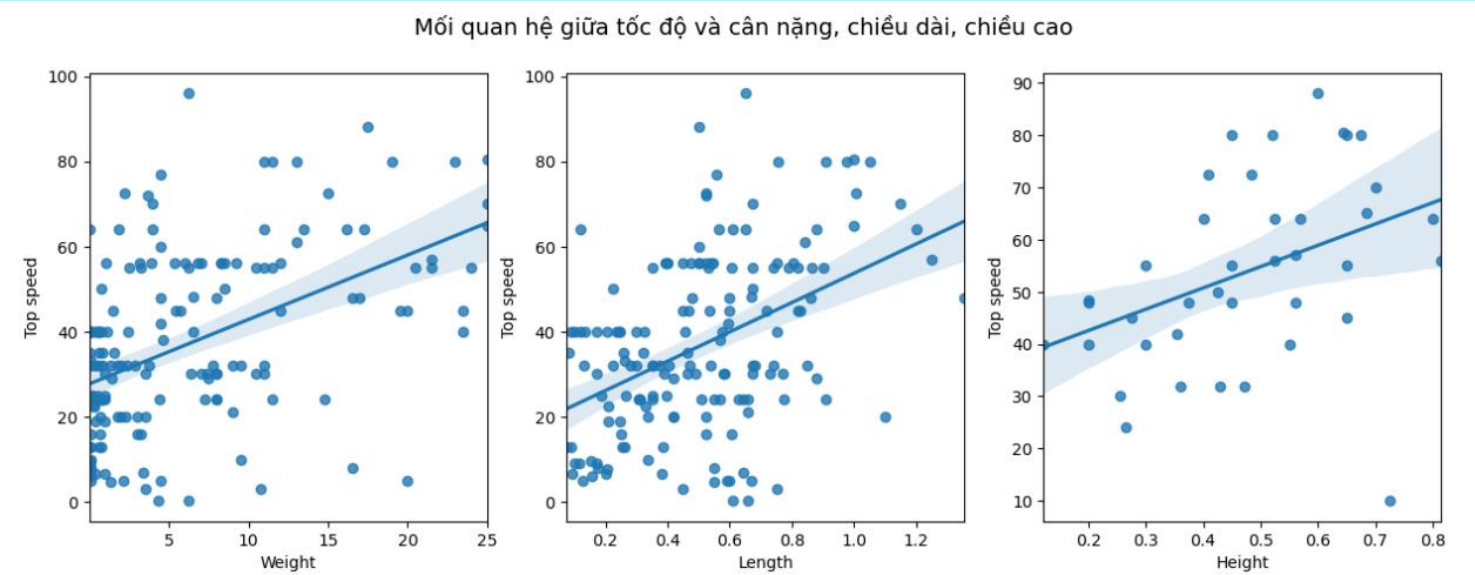
Bảng mức độ tương quan giữa các thông số trên:

- Độ tương quan giữa các thuộc tính 'Weight', "Length", 'Height' đối với 'Top speed' không quá cao, bé hơn 0.5.



5. Quan hệ giữa chiều dài, cân nặng, chiều cao với tốc độ

Biểu đồ thể hiện mối quan hệ giữa tốc độ với cân nặng, chiều dài, chiều cao để xem các điểm dữ liệu phân bố như thế nào:



5. Quan hệ giữa chiều dài, cân nặng, chiều cao với tốc độ

Biểu đồ thể hiện mối quan hệ giữa tốc độ với cân nặng, chiều dài, chiều cao để xem các điểm dữ liệu phân bố như thế nào:

- Có rất nhiều điểm dữ liệu nằm rời rạc và cách xa đường hồi quy -> có vẻ như các yếu tố cũng có mối liên hệ đến tốc độ của động vật đó nhưng không nhiều.
- Ở biểu đồ thứ 3 thì dữ liệu cũng khá ít nên chúng ta vẫn chưa thể kết luận gì nhiều về ảnh hưởng của chiều cao đến tốc độ của loài thuộc lớp thú.

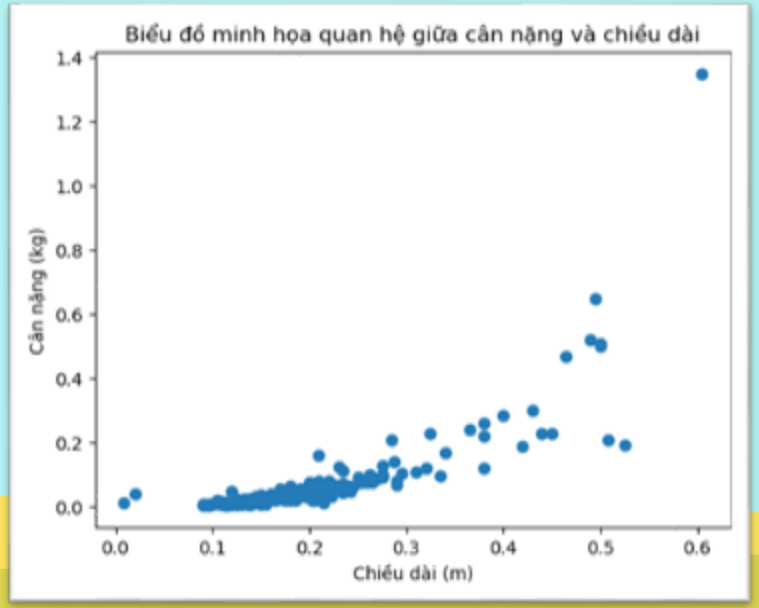
04

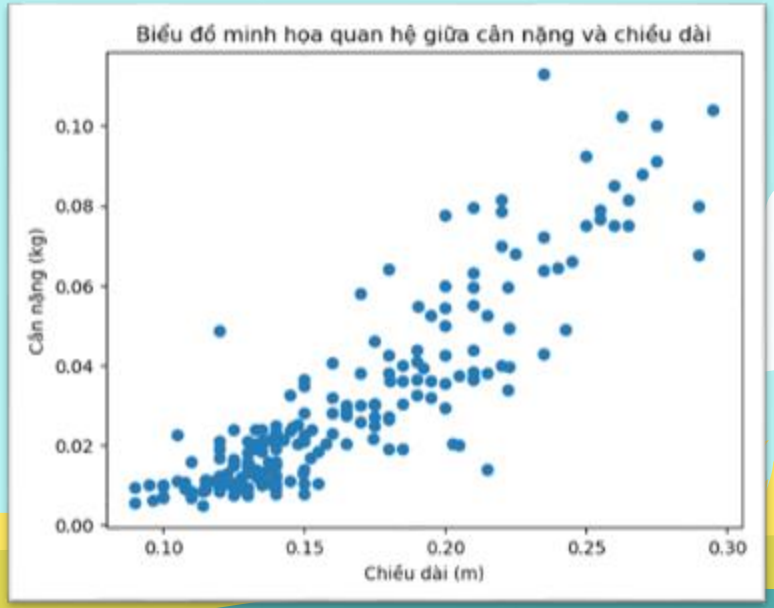
Mô hình và đánh giá

Xây dựng mô hình học máy và đánh giá mô hình đó

1. Tiền xử lý

Loại bỏ các giá trị nan và outlier

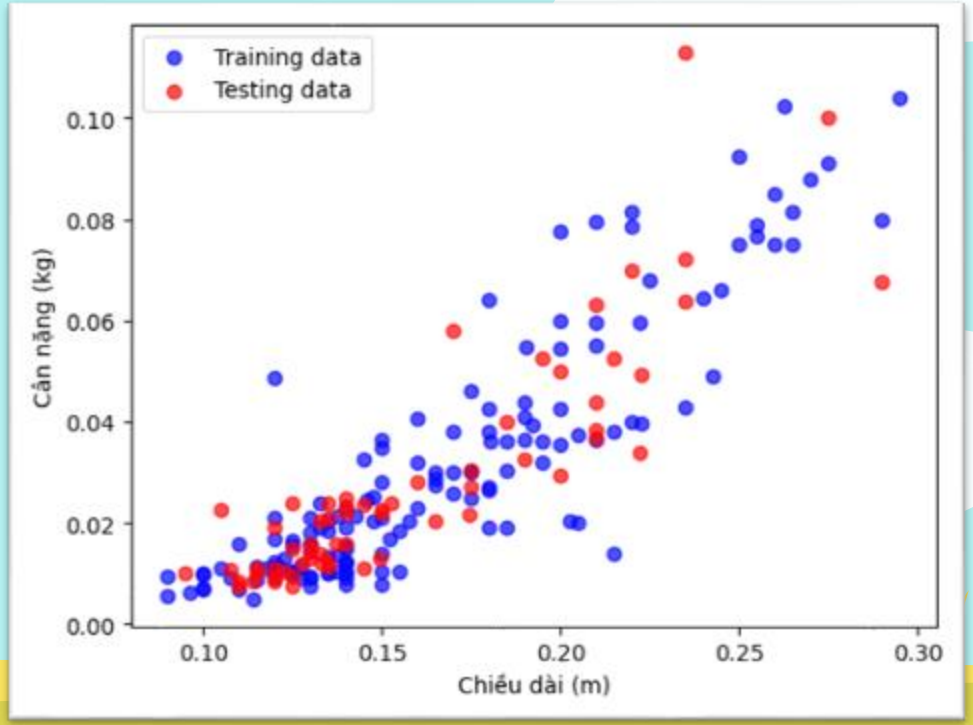




1. Tiền xử lý

Chuẩn bị dữ liệu

Chia tập dữ liệu ra tập train và tập test,để thực hiện việc huấn luyện trên tập train, predict trên tập test



2. Linear regression model

Huấn luyện

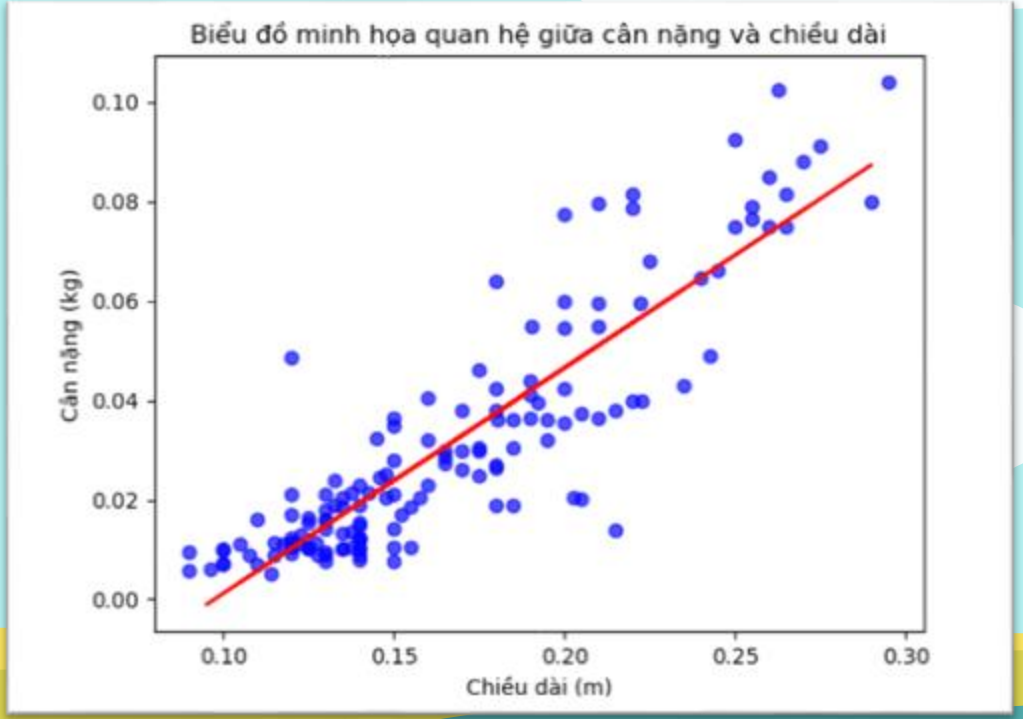
Thực hiện huấn luyện trên tập train và predict trên tập test để kiểm lại, rồi vẽ hình minh họa

Đánh giá

Dùng K-Fold cross-validication (k=5) rồi lấy mean để đánh giá

MSE: 0.0001274

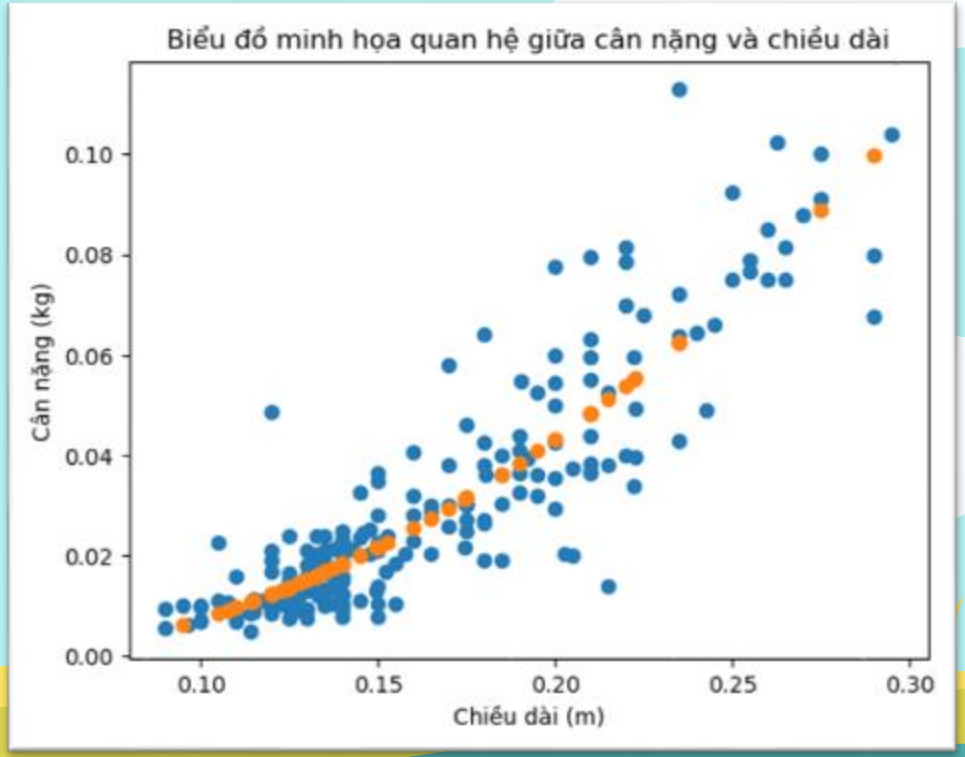
RMSE: 0.0111446



3. Polynomial regression model

Huấn luyện

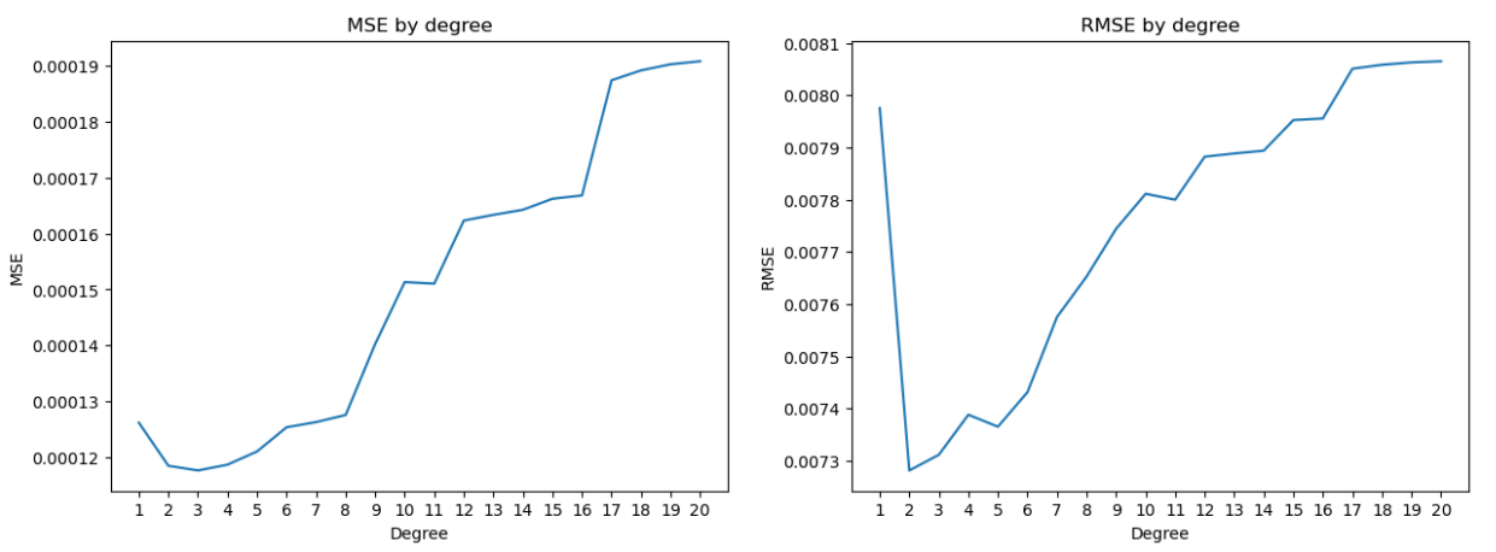
Thực hiện huấn luyện trên tập train và predict trên tập test để kiểm lại, rồi vẽ hình minh họa



3. Polynomial regression model

Tìm bậc tối ưu cho model

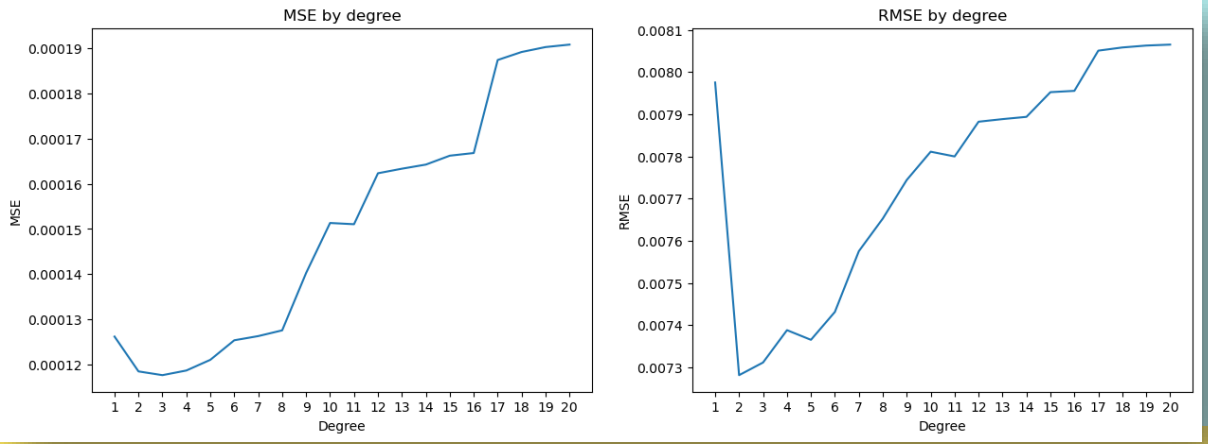
- Sử dụng phương pháp đánh giá Leave-one-out Cross-validation và chạy mô hình với bậc của đa thức từ 1 đến 20 để xem xu hướng sai số và cực tiểu sai số là bao nhiêu. Từ đó xác định được bậc tối ưu.
- Với bậc 2 và bậc 3 thì sai số MSE và RMSE sẽ thấp nhất.



3. Polynomial regression model

Đánh giá

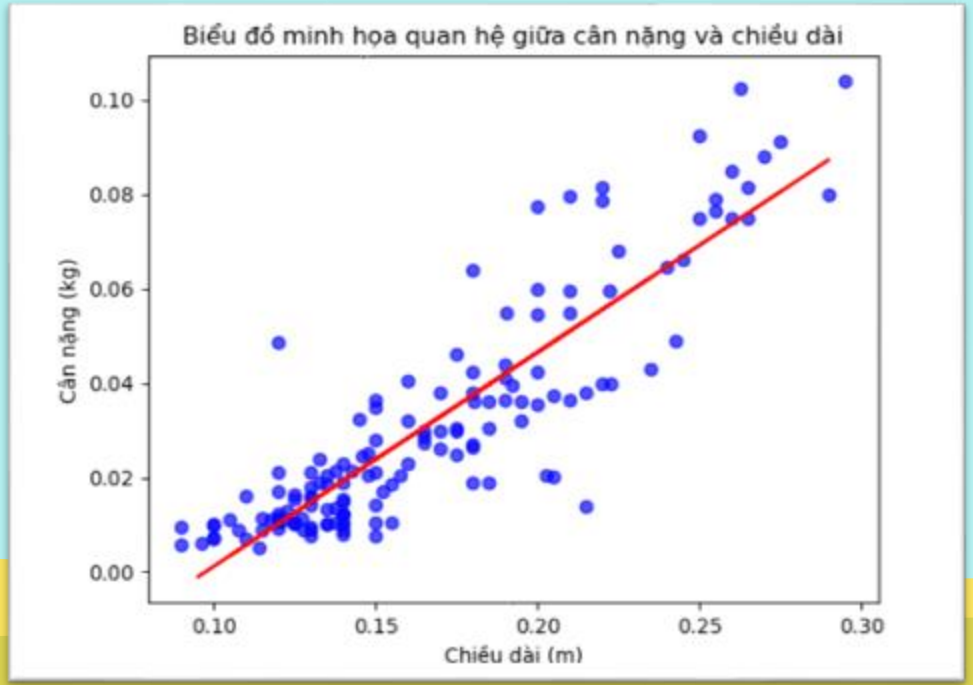
Khi bậc bằng 2 và 3 thì độ chính xác mô hình là tốt nhất.



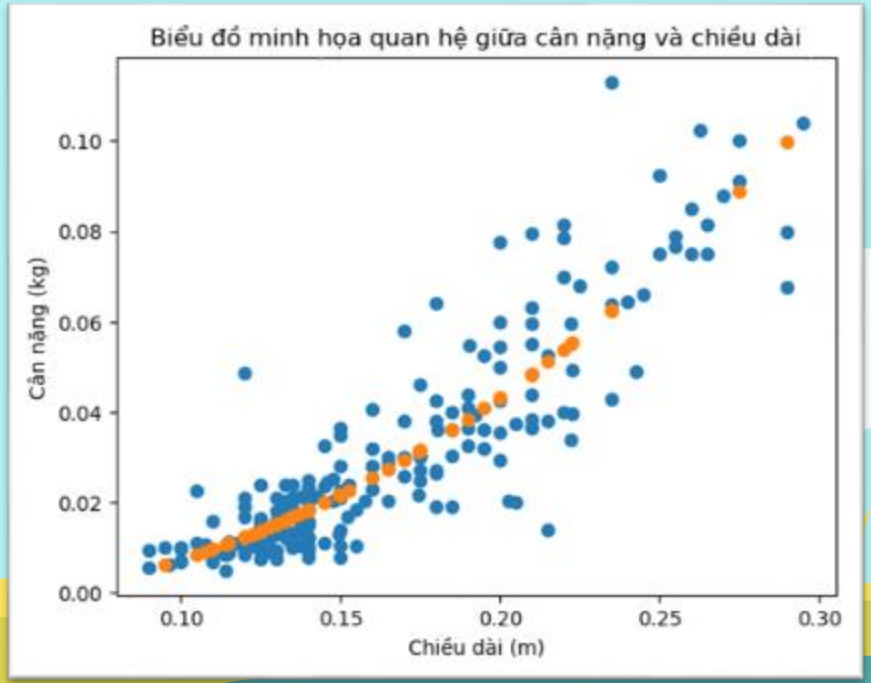
Bậc bằng 2:
MSE: 0.0001185
RMSE: 0.0072815
Bậc Bằng 3:
MSE: 0.0001177
RMSE: 0.0073115

4. So sánh 2 mô hình

Linear Regression



Polynomial Regression



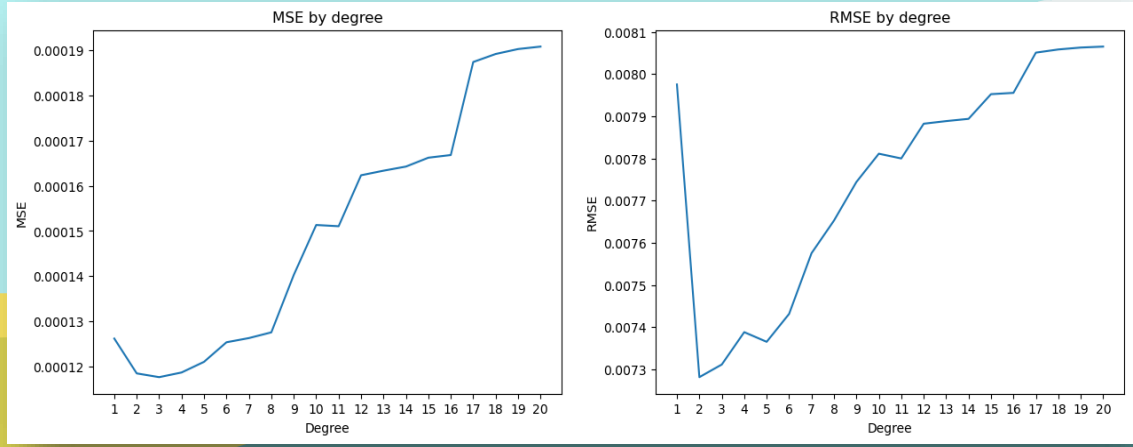
4. So sánh 2 mô hình

Linear Regression	Polynomial Regression
MSE: 0.0001274 RMSE: 0.0111446	Bậc bằng 2: MSE: 0.0001185 RMSE: 0.0072815 Bậc Bằng 3: MSE: 0.0001177 RMSE: 0.0073115

- Polynomial Regression nhất với bậc bằng 2 và 3 và tốt hơn Linear Regression (MSE).

- Polynomial Regression tốt hơn Linear Regression (RMSE) với mọi bậc nhỏ hơn 20.

Kết luận: mô hình 'Polynomial Regression' nhìn chung sẽ tốt hơn đối với bài toán và tập dữ liệu này.



Tài liệu tham khảo:

https://pandas.pydata.org/
https://matplotlib.org/
https://numpy.org/
http://seaborn.pydata.org/
sklearn.linear_model.LinearRegression — scikit-learn 1.2.0 documentation
sklearn.model_selection.cross_val_score — scikit-learn 1.2.0 documentation
Python | Linear Regression using sklearn - GeeksforGeeks
sklearn.preprocessing.PolynomialFeatures — scikit-learn 1.2.0 documentation
Các file notebook của Lab1, Lab2, Lab3
Slide bài giảng môn học Nhập môn khoa học dữ liệu.

THANK YOU