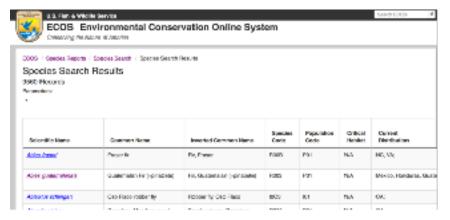
Assignment #2 Scraping and Scrubbing Data Simon Myunggun Seo (ms9144)

The data I was initially curious to see was the population of cute polar bears over the years. However, I could not find such integrated data that spans several decades.

Inspired from this idea, I started searching for data of other endangered species and came across a database called ECOS (Environmental Conservation Online System) hosted on the U.S. Fish & Wildlife Service website. This seemed reliable since it was maintained by a federal agency. I took the entire database (https://ecos.fws.gov/ecp0/reports/ad-hoc-species-report-input) by selecting all fields possible and I saved the webpage and mirrored it on my website (http://i6.cims.nyu.edu/~ms9144/external/Species_All.htm).



According to this database, there were at least 9,500 species that were ever considered to be fragile species in all kingdoms of organisms. This database has specific information about the habitat of each species and also records government documents so it could be useful for creating conservation campaigns for each region or state.

There are 17 fields of data for each species in this database:

Scientific Name, Common Name, Inverted Common Name, Species Code, Population Code, Critical Habitat, Current Distribution, Family, First Listed, Group, Lead Region, Federal Listing Status, Regions of Occurrence, Special Rules, U.S. or Foreign Listed, Vertebrate/Invertebrate/Plant, Where Listed

I wrote a Python script to scrap all the data exactly into a CSV file.

However, this data had a lot of useless information for me. The 17 fields include ecological, biological, taxonomical, and administrative data. Some of these were easy to figure out what they mean, but some (like species code: R00B) were of little meaning to

Scientific Herne	Goramon Harse	Inverted Common Hame	Openies Oods	Papelation Sode	Critical appropria	CurvedOstribution	Fornity	Soucies Group	Lead Fo	Federal Listing 6to	Regions of Goran	Special Rules	Vertebestoring
Vicina traced	Frants	Fit; Fisasor	RDOR	P91	NA.	NC: WV	Pleacone	Countries and Cytoses		Not Listed	4.6	167.	ρ
district grantematicanis	Destruction Fir (cylesteric)	Fig.G. externalism (optimizate)	MDCA.	P91	NA.	Masters, Heresteren, Con-	Piercon	Continue and Cyronic	Permiyo	Threateneri	164	14/5	js.
Abbutus soldingerii	this Flaco sobber fly	Follow fly Dec Place	1050	101	KA.	ek .	Asilder	hects		Mat Dated		943.	1
Korona apara	Mareinan Planting sanchystera	Sand-versera, Ramenay Mesdovs	GOOM	PM	14/4	KA;	Нустаўгоская	Rowering Franks		Not based		NA.	ρ
Boranis armoghila	Milescolene Konci Verbrees	Reviews Nationalisms Basel	ann	P91	NA.		Hydaghorean	Florening Floren.	5	Net Listed	164	140.	js.
Foranio emmoshila nen	Ne common memo	No common name	G3U4	P91	NN.		Hydraghoosso	Flowering Florito	- 5	Het Listed	194.	140	la .
Porone arrangales	(Deserved) sand-vertiens	Sand-versena, (ernamed)	COMM	P28	Ps/No	wc	Hydaginsoxae	Revening Plants		NEC LIENCE		MA.	P
Expression .	julimented junto-vertiene	Sand-remena, jumamed)	000%	PR	PuN.	NHC.	Нустарігескан	Revening Plants	- 2	Not Linked	2	16/4	ρ
Abronia macrosegue	Large fruited send-verberes	Serdi-rerbens, large-huted	3490	P94	N/A	TRUBBLE ALCTRO	Hydaghsoms	Revering Frants	i i	Endengend		64/Ju	p
Foronia turbinata	[Jenomed] sand-verbens	Sand-vertions, [unnamed]	COF.	P91	NA.	W/S	Nydaginosao	Revering Plants	1	Not Listed	8	167.	ρ
Koronis unicellata acutalista	Mass-purprises divertiess	Sand-versions, con-purple	CENTY	PST	Pc%	WA Positify sales in	Nychaginicasa	However, Franks	- 1	Not Unled	1	160	ja .
Abronia umbeliale brevillere	Pricerd-vectors	Sandi-rambena, pink	GSAS:	198	N/M	KA, OR:	Nydaginsbass	Revening Flants		Not Listed	1.8	14.5	Į1
Aburillor enemittypetalum.	Me common marries	No common name	0154	P94	47.0504	PC U.S.A. (18)	Makwoose	Revering Flants		Endergered	1	140.	p
Author halosonum	Feedoprio	Redopelo	G008	P91	16%	PR.Vt.	Makraossa	Revering Rants	- 4	Mat Listed	4	NA.	ρ
Kodomeralesi	Kronseste	Postoriula	Q186	PST	Pc%	R. USA (H)	Matracese	Howevey Hants	- 1	arcangend	1	NA.	ja .
doubles pariabili	Me common mans	No common rasma	624	P94	N/h	42:	Makwees	Revering Flants	è	Hint Listed		147.	P
Abution sandwicened	No common memo	No common name	019.	P21	17.05(a)	Pt: U.S.A. (18)	Malvecoop	Flowering Floats	1	Endergend	1	140.	l _e
Kouthon respensarior	No common same	No common name	CORNE	PZE	Pc%	MK, Vt. Press bly solved	Matracese	Howevery Franks	- 6	Not baled	4	NA.	je.
Acuda Koula	Rade	Roule	G000	P91	N/N	R:	Pathecover	Reverte Fluids	1	Net Dated	1	NA.	P
Acresis solicas	Libral	Liliensi	01584	P94	17,9608	BLUGALER	РОМИНИ	Revering Plants	1	Endangend	1	MA.	P
Analysiaen susanne	[Limoved] ince itug	Laco (sig. [.nearesc]	KMM	101	NA.	AR;	Tingicin	Projects		Not Littled	4	167.	1
Aueritracyalops antenidoreis	Capragnosi	Dispreparal	KD66	ICI	NN.		Capaporio	Deniscore	HA	Nat Listed	104	160	1
Acarthometropus pecetosics	Prostoelos Siver martis	Maylly Recatosics River	1000	101	14%	ILW:	GgMosuridae	Praects		Not Usted		14.5.	1

me. Excluding fields that I did not have an intuitive understanding of, I decided to keep the following 8 fields:

Scientific Name, Common Name, Current Distribution, Family, Group, Federal Listing Status, Regions of Occurrence, Vertebrate/Invertebrate/Plant

Hopefully someone with more expertise can give an explanation for what each fields exactly describe.

As I could not simply visualize this qualitative data easily using graphs, I sorted each column and scrolled through, focusing on the top and bottom of the table for anomalies. There weren't significant anomalies although I found that some cells needed a small cleanup. Each field noted unavailable data in different ways.

The "Current Distribution" field, which describes the names of countries or states that each species is observed, left unavailable data empty.

The "Regions of Occurrence" field, which describes broad geographical regions in the US where each species is observed, marked it as "NA" which made me confused whether it signified North America or Not Applicable.

The "Common Name" field, which describes common, as opposed to academic, names of species, marked it as "No common name".

Some fields (that I decided to remove) marked it as "N/A".

I decided to reiterate through the data and unify everything as "N/A". However, I left this one alone: "Federal listing status" which describes whether the species is listed as endangered/threatened according to US federal data marked it as "Not Listed".

Substille Home	Coremon Home	Convertibilitation	Forth	Species Group	Perform Library Status	Region	of Device retries	Metabole/Investeinste/Plant
Amphiano photoler	Onertains Ansahrlana	AL PL. GA LIS.	Anglitonian	Analdian	Sinche Review in the Co			. 4
Eleutreeodactylus imperi	Golden coqui	PR: Endertic to Puerto Ricz, U.S.A. (PR)	Luptochristone	Amphibians	Threstered			W.
Electrocodedylas kanartyerol	Ularano Cespi	PR-USA (PR)	Liptorarylow	Arrohibans	Brdangered			¥
Electrocodectylus are dise	Mothed cocul	PR.Vb	Laptochristicae	Amphibians	Ret Livted			W.
Electronoductylus activantal	Virgin Islanda cogal	Wigh Munds	Luctockristick	Amphibians	Ret Uirled	NA.		W.
Eleutreroductylus karladeraidd	Weo-foreid cogsi	PR.Vb	Luctockraticke	Amphibians	First Livred			W.
Breating exchectofall Kauberi	Large-blotched essains	DA;	Piethodort day	Amphibians	Ret Uirled		1	W .
Executive excharcholistic processor	Vellow-biotched aneatine	DA;	Piethodorádae	Anghibiana	Rot Listed		1	W .
Rang despicell	California red-Inograf Iraq	DA, U.S.A. (DA, Minim	Raridas	Analitians	Tirestored			W
Rano amoista capita	Carolina oranifer ling	9A, 10, 90,	Raridas	Analitians	Brut Listed			
Fano consocior	Carcadas log	DA, OR, WA	Raridas	Anshibere	Sincle Review in the Co.	1, 8		v
Fano déricaloresis	Chircat na leopard frog	42. HA, 3.9.A. (MI), A2), Market	Raridas	Amphibians	Tirestored			W
Earny Inderton in to	Columbia spelled hop	D, No. CR, . Sesed on propriets and genetic obsorber salon. Date Fusik, M.G., C.A. Posti, H.M. Droheim, M.J. Adams, T.C. Mullins, and S.		Architers	But Usled	1, 8		w
Fano latelysotria	Galumble spetted frog	D. Oft	Floridos	Amphibians	Not Listed			*
Fare:	Duskmoter freg	NY:	Floridos	Anchibiana	Nat Listed		1	¥
Famo arcolate sorces	Dualry convision ring	ALFL LA HS	Floridos	Anchibiana	Nat Listed			· *
Farm couces	dask, godner freg	US USA (MS. IA AL)	Ranidos	Amphibians	Endangered			W
Farm chalcoure	Florida beg flog	R.:	Ranidos	Amphibians	First Linted			W
Fano amouta tosapes	Fitnish convetica rang	R.,GA	Boridos	Anchibians	Not Littled			
Fann beylii	Feotolitystess- egged hop	DA, ORI	Boridos	Anchibians	Sander Review in the Co.	1, 8		w
Commun goliath	Gatilatin Freig	Sance, Equatorial Garrier, Comerces	Boridos	Anchibians	Threatened	1975		w
Lithatester sapito	Gapher Freq	Ra	Foridos	Anahibians	Under Review in the Co.			. *
Pocudanto Illevenia	Minute of turns frequency	AR IL NO:	Pylidon	Amultiblero	Single Review in the Car	\$ ₁ t		4
Discoplinates stationales	forced painted floor	bread	Disectionists	Amultiblero	Briangered	165		4
Fami successive also	Leviand record in San Falce	AZ-CA MM, UE	Paristra	Amultiklaru	Bret Libbed	1.1.8		4
Family recoverse	Hourisis yellow leyers from	DAUSA OR ON	Paristra	Amultiklaru	Entergrad			¥
Farst Property	Hourston yellow bropps from	04	Beriden	Arabiliary	Entergreed			¥

Because "Not Listed" is a specific status, it is not the same as the above cases of unavailable data which have no existing or observed data.

Now the data is a bit cleaner, but I thought I could make it more useful at a personal level. I decided to check if each species were observable within NY or the Northeast Region of the US (referred to as region #5) and save that as data. I wrote a filter() function that checks whether the given field includes a certain word and then saves the result as a Boolean value in a new data field.

Scientific Name	Commen Samo	Current Distribution	Early	Species Group	Fodoral Litting Status		Mortebross/im	n HY in Gurror (Cistellastion	6 in Poploso of Government
					_	_	-		
Lampolitio salnocivori		AR: Little Red Fliver watershed, Arkensess; U.S.A. (AR) Ender		Chra	Chdangered		1	false	Caine
Popagene elterame	Centa Lucia pagagona	6A:	Latriscese	Flowering Plants	Not Listed		P.	Fe/se-	Fo/se
Popegene douglas i parvirlors	Small-Roward, Douglar' page	CA:	Latitioner	Floreting Plants	Not Ushed		P	Palor	Paler
seiria medeoloideo	čeroli whorled pogenia	OT DE GALL, MAINE MILMOING, MH, MJ, NY, GH, PN, N	Grafrideopea	Florating Floria	Trentance	2.4.6	P	Tue:	Tra
Mayora Indones	Lower or propositi	BUUSA (FL)	Programme	Forming Plans	Findingswort	4	μ	Forte	Forter
Yolygina kinathi	jick bestättigen	REUSA (R)	Pringracese	Flowning Place	Endangered	4	P	Farse	Earte:
Yolygonum friownenii	Scots Valley Roygonum	OA, U.S.A. (DA), Honely of the Ony of Scots Valley.	Pritygonicase	Flowering Plants	Endangered		P	False	False
Stammilia nimbata	Dwsfipolypody	HO:	Gramm Michigene	Ferns and Alfee	Not Listed	4	p	False	False
Droscphile terolomeis	Land considerly	Ht. Possibly extract not reported since 1965	Drosophildisa	Insects	Not Usrael	1	1	Polye	Paler
Jedoro reoliosifolia	Perduty	AL ARIGA, VO. MS, VC. SQ, U.S.A. (SC. HC. MS, NO. LA.	Lowerowe	Floralis Plans	Endangered	2.4	r	Folum	Folor
iongeliante konnectionale	For whomat juicovery the pend	ni	Lymmetrian	Sivolit	Sectional		1	Faite	Entre
Sagritude statement	Trickshel pordscar	UI; Selevectic he extinct.	Lynnaedae	2hols	Not used		1	Palse	Palse
Base antivale	Pandigica	R.,	Lauraceae	Floreing Plants	Not Ushed	4	P	False	False
Potamogeton confervelides	(Linearmed) grondwood	DE MALME, NIL NO, NH, NIL NY, RALSO, VE WIT	Petamegatenace	Florening Plants	Not Ukhadi	1 4 5	P	Tur	Trans.
Potamogoton ogdenii	(Linearmod) grandwood	MK-HCYT:	Petamegatenoso	Florating Floria	Not Listed		P	Tue:	Tue .
Potomogaton Karldonia	Florida portilental	n:	Petarregalenos	Floresing Florin	Under Review In the Ger		Pi	Polym	Polon
Salamagatan kilki	HITE SOMMONE	C1, MA, M1, H2, CH, P5, V1; Clower (SWI)	Peterregetenson	Horning Plans	Set Med	2, 6.	p	No.	Yes
Otamogeton obydoserous	Time Agrya subsemp Principles	15; #86 (IS)	Pitaragnasos	Provening Places	Endangered	2	μ	Nation	Note
							-		