Milestone 1- Project Proposal

1. Group member:

Xiangchen Guan: xcguan98@upenn.edu
Minzheng Zhang: minzheng@seas.upenn.edu
Zijian Xiao: zijianx@seas.upenn.edu
Peiran Xu: peiranxu@seas.upenn.edu

2. Description:

We would like to develop an application which is relevant to wild animals and national parks. In our application, the user will be able to find the animal species inside each national park, and for each kind of wild animal, which national parks do they live in right now. The user will also be able to see the statistics such as the relationship between the number of endangered animals and the biodiversity of the national parks, or the relationship between animal nativeness and animal abundance of the park. The user could also get a brief introduction and pictures of species.

3. Dataset:

   a. Wild Animal Trade

   Link:https://www.kaggle.com/datasets/cites/cites-wildlife-trade-database

   Description: This dataset contains records on every international import or export conducted with species from the CITES lists in 2016.

relevant size statistic	6.2MB, 67161 rows, 16 attributes
summary statistic	

	Year	App.	Taxon	Class	Order	Family	Genus	Importer	Exporter	Origin
count	67161.000000	67161	67161	46937	67104	66700	65702	67090	66588	25643
unique	NaN	4	6382	16	101	252	1340	216	211	179
top	NaN	II	Crocodylus niloticus	Reptilia	Orchidales	Orchidaceae	Crocodylus	US	NL	ID
freq	NaN	59253	2417	18430	9973	9973	4408	9722	7201	2974
mean	2016.002293	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
std	0.047831	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
min	2016.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
25%	2016.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
50%	2016.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
75%	2016.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
max	2017.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Exporter reported quantity	Term	Unit	Purpose	Source
4.402100e+04	67161	6402	61102	66617
NaN	83	13	12	10
NaN	live	kg	T	W
NaN	33862	3643	47081	21590
4.443878e+03	NaN	NaN	NaN	NaN
1.573794e+05	NaN	NaN	NaN	NaN
2.000000e-06	NaN	NaN	NaN	NaN
2.000000e+00	NaN	NaN	NaN	NaN
1.200000e+01	NaN	NaN	NaN	NaN
8.200000e+01	NaN	NaN	NaN	NaN
2.154362e+07	NaN	NaN	NaN	NaN

   b. Biodiversity in National Parks

link: https://www.kaggle.com/datasets/nationalparkservice/park-biodiversity?select=parks.csv
Description: a database of animal and plant species identified in individual national parks and verified by evidence — observations, vouchers, or reports that document the presence of a species in a park.

relevant size statistic	17.5MB, 119248 rows, 14 attributes
summary statistic	

	Species ID	Park Name	Category	Order	Family	Scientific Name	Common Names
count	119248	119248	119248	117776	117736	119248	119248
unique	119248	56	14	554	2332	46022	35826
top	ACAD-1000	Great Smoky Mountains National Park	Vascular Plant	Poales	Asteraceae	Falco peregrinus	None
freq	1	6623	65221	11453	8843	56	27147

Record Status	Occurrence	Nativeness	Abundance	Seasonality	Conservation Status	Unnamed: 13
119248	99106	94203	76306	20157	4718	5
54	7	5	8	24	11	3
Approved	Present	Native	Unknown	Breeder	Species of Concern	Threatened
86254	83278	75950	28119	12214	3843	2

c. (optional) Wikipedia
Fetch some brief introduction and pictures of the species from the web.

4. Queries:
The two dataset are joined on species' Scientific Name
   a. We would like to join Biodiversity National Park dataset with Wild Animal Trade dataset to see which national parks the imported animals currently live in
   b. We would like to show if the total number of endangered animals is associated with the biodiversity of a national park by calculating the Shannon diversity index from the abundance of all species/class in the park.
   c. We would like to show the number of national parks that have rare animals exported by each country, in other words, for each country, we want to know how many national parks have rare species exported by this country. We do this by joining Biodiversity in National Parks and Wild Animal Trade on which abundance is rare and count the distinct number of national parks with respect to the group of rare species for each country.
   d. We would like to show the number of invasive species(not native) that each country exports to each national park by selecting the non-native species from each national park in Biodiversity in National Parks and joining into Wild Animal Trade to see which export countries these species belong to and count the number of non-native species that each country exports to each national park.
   e. We would like to show where the hotspot origins of imported animals are in different National Parks by aggregating the origin country in the trading where the importer is US.

5. Github Link: https://github.com/estherxpr/CIS550_final_project