Project Idea Report

Jian Di, Deyang Li, Ding Ning, Chen Sun

September 22, 2019

***Part* 1.** *Idea*

The goal of this project is to discover the endangered animals and the factors that possibly threaten them, for example, the habitat loss, the climate change and the disease.[1] Our common sense tells us that the increasing human population and the partial economic growth may deprive the habitats of an animal, and the climate change results in the ecosystem shifts. However, does the data tell us the same stories as what we assume? This project will transform the data acquired from the online sources associated with endangered animals, into a format suitable for analysis and visualization. Then we want to use the transformed dataset to see if the endangered animals are threatened by these assumed factors, making them understandable for our audience.

***Part* 2.** *Source*

1) The preliminary dataset includes the animal data from A-Z Animals (https://a-z-animals.com/). We used R to conduct the web scrapping and stored the data in a CSV file. The dataset consists of 593 observations and 33 variables.

2) Why A-Z Animals?

   a) The web site provides the information of the 593 well-known animals, such as the name, the location, the biological information, the conservation status, the estimated population size, and the biggest threat. The information is non-technical and easily understood.

[1] National Wildlife Federation. (n.d.). *Threats to wildlife*. Retrieved from https://www.nwf.org/Educational-Resources/Wildlife-Guide/Threats-to-Wildlife.

b) The web site has well organized the information of each animal. The data is structured.

c) The web site provides the pictures of the animals.

d) The endangered animals are clearly labeled and some have mentioned their biggest threat.

e) Most information is permitted to be used freely.

3) Problems of the A-Z Animals dataset.

a) The information is incomplete, so there are many "NAs" in our preliminary dataset. The other sources may be used to fill in these empty cells.

b) Some variables are irrelevant. We may not want the information that do not explicitly reflect the endangerment, such as the family, the scientific name and the color. Filling in these empty cells will take extra time.

c) Some observations are irrelevant, namely those of the unendangered animals.

d) The biggest threat is described in a few words, which are unable to be directly used. With these texts, we need the structured data from the other sources.