Information Visualization
Checkpoint I

Checkpoint I: Project Proposal

Group: 34
Date: <2025/09/14>

Problem Domain

With this project, we aim to study different crocodile species across many different environments. The analysis of this information will help educate and elucidate how the crocodile species differ and how those differences are in any way related to their environments. To achieve this, we collect comprehensive data on the physical characteristics and on the habitats of each species in the form of measured observations.

The study of crocodile species and their relation to the environment is important because crocodiles are apex predators and play significant roles in maintaining the balance in their ecosystems. This means that their overall health can provide information about the health of their habitats.

Task Abstraction

There are many questions that this project aims to answer, each pertaining to information that aims to educate or elucidate how each crocodile species differs and if those differences are related to their environments. The questions we aim to answer are as follows:

- What is the distribution of species found in each country? This is a distribution question. This question can be phrased, abstractly, as: what is the distribution of X in Y? How many of each species can be found in India?
- Is there a correlation between a crocodile's weight, length and gender? Are male crocodiles longer and heavier than female crocodiles? This is a correlation question because we are looking into the attributes weight, length and gender, and trying to find out if there are any links between them. This question can be phrased, abstractly, as: is there a correlation between attributes X and Y and Z?
- Are crocodiles being observed now more than there were a number of years ago? This is a trend question. This question can be phrased, abstractly, as: does X increase with Y? Are there more observations now than there were 2 years ago?
- What is the distribution of ages of crocodiles in countries where they are in more vulnerable conservation statuses? This is a distribution question. This question can be phrased, abstractly, as: what is the distribution of X in Y where Z?
- What is the distribution of species found in each habitat? This is a distribution question. This question can be phrased, abstractly, as: what is the distribution of X in Y? How many of each species can be found in swamps?

- Is there a correlation between weight, length and habitat? This is a correlation question, since we aim to find out if there is a link between several different attributes. This question can be phrased, abstractly, as: is there a correlation between attributes X and Y and Z? Can heavier and longer crocodiles be found more in swamps?

Data Abstraction

The dataset used in this project is called "Global Crocodile Species Dataset" which can be found on Kaggle[1]. It is a static dataset as can be seen by the fact that Kaggle states that updates for the dataset are not expected and the update history of the dataset shows that its only update (as described by Kaggle) was its creation date (27/08/25). The dataset is composed of 15 attributes and 1000 items and, while there are some attributes that can be cleaned, we considered the size of the dataset appropriate for the tasks at hand.

To acquire the dataset, all one needs to do is create a Kaggle account and download the dataset, which already comes in the CSV format inside its compressed file. As stated before, the initial dataset has 15 attributes, but we do not need to work with those 15, some can be cleaned. We therefore remove the following attributes:

- Observation ID: the unique ID (labelling each item from 1-1000) is not useful information for our purposes, nor is it relevant for any of the other attributes (if ID 1 were 67 instead, it would make no difference for the data).
- Scientific Name: The scientific name attribute will always change accordingly with the Common Name attribute. It is therefore considered redundant information in the context of this project.
- Family: the family attribute has a single unique value across all items in the dataset ("Crocodylidae"). Since there is no variation in its value, it provides no useful information, at least in the context of this project.
- Genus: the genus attribute is not relevant to any of the proposed questions.
- Observer Name: This information is not considered relevant for any of the questions we have proposed for our project, and it is not considered data pertaining to the crocodiles themselves.
- Notes: This attribute is also not considered relevant for any questions we have proposed. It is also highly descriptive data that is unique for each item and varies too much.

We decided to keep the remaining 9 attributes, for the following reasons:

- Common Name: nominal attribute type since it is composed of names, which can't be ordered and only compared if equal or not. This is the name that most people will know the species by rather than its scientific name.
- Observed Length (m): varies between 0.14 and 6.12, it is ordered, can be added together, and it's possible to establish ratios as well, making it an ordered quantitative ratio attribute type. This is the measured length, from tail to snout, of the observed unit.

- Observed Weight (kg): varies between 4.4 and 1139.7, it is ordered, can be added together, and it's possible to establish ratios as well, making it, similar to the previous attribute, an ordered quantitative ratio attribute type. This is the measured weight (in kilograms) of the observed unit.
- Age Class: ordinal attribute type comprising 4 different age classes: Hatchling, Juvenile, Subadult, Adult; in respective order from youngest to oldest. This is an ordinal attribute type because although we can establish an order, we can't quantify the interval of change between the different categories. This is a descriptor of a rough estimation of the crocodile's age.
- Sex: has 3 different categories: Male, Female and Unknown. There is no order between them so it's a nominal attribute type. This is the observed sex of the crocodile, and it is important to note that some crocodiles did not have their sexes observed.
- Date of Observation: ordered quantitative continuous attribute. It is a date in the form of DD/MM/YYYY.
- Country/Region: composed of names, making it a nominal attribute. It tells us in what country or region the observation took place.
- Habitat Type: also composed of names, making it a nominal attribute. It is a descriptor of the nature of the habitat in which the observation took place.
- Conservation Status: attribute with 5 categories: Data Deficient, Least Concern, Vulnerable, Endangered and Critically Endangered. There is an order between the categories, but the extent of the interval between them can't be quantified, making it an ordinal attribute. It is a rough descriptor of the extinction status of said species.

Each item represents an observation of a crocodile. All attributes are fairly self-explanatory, only needing to highlight that a Sex of unknown means the observation could not identify the sex of the crocodile and that Data deficient in conservation status means there is not enough data pertaining to the current state of conservation of the crocodile species. The dataset is presented in a tabular format and, while it does contain information pertaining to geography, the dataset is still tabular.

As noted before, there are a few missing values, namely the "Unknown" and "Data Deficient" values in the "Sex" and "Conservation Status" respectively. These missing values are significant in quantity in relation to the size of the dataset (35.4% have Unknown sex, 11.5% have Data Deficient conservation status). Our solution to these missing values was to consider "Unknown" as a sentinel value, while considering "Data Deficient" a non-missing value. While "Data Deficient" itself does mean that there is not enough information to correctly categorize a species, the "Data Deficient" category itself and its name provide important information on the species as well as its region and habitat - "Regions predicted to contain large numbers of threatened DD [Data Deficient] species are already conservation priorities"[2]. Therefore we have opted for treating "Data Deficient" as its own value, just like any Conservation Status value. All of the questions can be answered without derived attributes as can be seen in the following "Mapping" section.

Mapping

- What is the distribution of species found in each country?
    - We can look at the attribute “Country” and use it as a filter for the attribute “Common Name”.
- Is there a correlation between a crocodile’s weight, length and gender? Are male crocodiles longer and heavier than female crocodiles?
    - We will use the attribute “Sex” and we will compare the items using the attributes “Observed Length” and “Observed Weight” across the different sexes to see if they are correlated.
- Are crocodiles being observed now more than there were a number of years ago?
    - We can use the attribute “Date of Observation” and compare it across time, going from 2005 to the present.
- What is the distribution of ages of crocodiles in places where they are in more vulnerable conservation statuses?
    - We can filter the attribute “Age Class” using “Vulnerable” as the filter for the attribute “Conservation Status”.
- What is the distribution of species found in each habitat?
    - We can look at the attribute “Habitat Type” and use it as a filter for the “Common Name” attribute.
- Is there a correlation between weight, length and habitat?
    - We will use the attribute “Habitat Type” and we will compare the items using the attributes “Observed Length” and “Observed Weight” across the multiple habitats to see if they are correlated.

References

1 - https://www.kaggle.com/datasets/zadafiyabhrami/global-crocodile-species-dataset

2 - Bland, Lucie M., et al. “Predicting the Conservation Status of Data-Deficient Species.” *Conservation Biology*, vol. 29, no. 1, 13 Aug. 2014, pp. 250–259, https://doi.org/10.1111/cobi.12372.