**Data Wrangling**

**Introduction**

Step **Data Wrangling** is the process that includes cleaning and storing data generated from the 5 websites mentioned in data scraping part. Data cleaning combines different data frame into one single file, filtering NA data, adding new columns and group rows together. The techniques which are used in this step contain R *Tidyverse, skimr and magrittr* packages and Julia *Pkg, CSV and DataFrames* packages. The summary is showed in the table below.

**Wrangling with R Programming**

Table 1- Wrangling Data Info

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Order | Data | Purpose | Techniques | Data Frame |
| 1 | A – Z animals data frame:  593 rows, 61 columns, 1 file | 1. Put two groups of data into one data frame  2. Select the columns and rows. | *Tidyverse,readr,skimr,plyr,lubridate and devtools)* | A – Z endangered animals data frame |
| 2 | A – Z animals image data frame:  592 images  592 names and links, 1 file |
| 3 | Endangered mammal animals data frame:  1329 rows, 10 columns, 11 files | 1. Combine endangered mammal animals in 11 areas/files into one single file  2. Clean the data and edit the data frame for the future analysis | *tidyverse, skimr, magrittr* | Tidy endangered mammal animals data frame:  1328 rows, 8 columns, 1 file |
| 4 | [Weather](https://www.ncdc.noaa.gov/cag/global/time-series/hawaiianRegion/land_ocean/ann/12/1880-2019) data:  14 files | Select the data related to our project |  | [Weather](https://www.ncdc.noaa.gov/cag/global/time-series/hawaiianRegion/land_ocean/ann/12/1880-2019) data frame:  1 file |
| 5 | Population data | Select the data related to our project |  | Population data frame:  1 file |

**Endangered mammal animals data**

The process of the endangered mammal animals data involves the functions of read\_csv, glimpse, skim, kable, mutate, group\_by, tally, summarise and sum.

We create a function (create.Area) which can add a new column named Area. Its value is based on another column “Endangered Animals of Area”. We extract the following part of the sub sentence “Endangered Mammals of ” and then use regular expression [^a-zA-Z0-9 -] to remove all non-numeric and non-alphabet characters before saving them as one single filed named “new-animals\_all.csv”. We load the data from new-animals\_all.csv into R for checking the types of the columns and the missing values, and output a quick summary of the dataset. After changing the data of Hawaii and Central America into North America and the data of Australia into the Oceans and Island Nation because of the classification of seven continents, we deleted the irrelevant variable named "Other Names and/or Listed subspecies" because it has lots of missing data. We also filter the only NA appearing in the “Status/Date Listed as Endangered” column because this information is unavailable on the website.

We use the mutate function to classify "T-" into "NT :Near Threatened" based on the definition of "T = threatened : a species likely to become endangered within the foreseeable future throughout all or a significant portion of its range” (<https://www.fws.gov/endangered/about/listing-status-codes.html>) and “NT :Near Threatened” from "The IUCN Red List Categories define the extinction risk of species assessed into nine categories" (<https://www.iucn.org/resources/conservation-tools/iucn-red-list-threatened-species>). According to the result of the group\_by function for the “Status” and “Area” columns, we use the tally function to calculate the total number of endangered animals in each areas for different status.

**Wrangling with Julia for endangered mammal animals data**

We create a new column area with the create\_Area(x) function. In this function, we use regular expression to match the part after "Endangered Mammals of " and numeric and alphabet characters. And then we use the CSV.read function to read all 11 endangered mammal animal files for 11 different areas and add one new column named “Area” with the insert! function. We also use the SubString function to make a index and copy the selected part of the original column of "Status" to the final data frame. In the end, we use the CSV.write function to get the csv document named “new-animals\_all\_Julia.csv”.