Planet Zoo Trivia

Gathering the Data

Sergei Dakov

additional packages:

```
library(rvest)
```

used to scrape data from the web

```
library(stringr)
```

used to tidily process strings

```
library(dplyr)
```

used to tidy multiple operatons into pipes

First, using the rvest browser, go tothe page that holds the master table of all available animal types wiith links to their individual pages

```
current_session <- session('https://planetzoo.fandom.com/wiki/List_of_Animals')
```

create a table to contain all needed categories. (categories were chosen based ondata availability on the site)

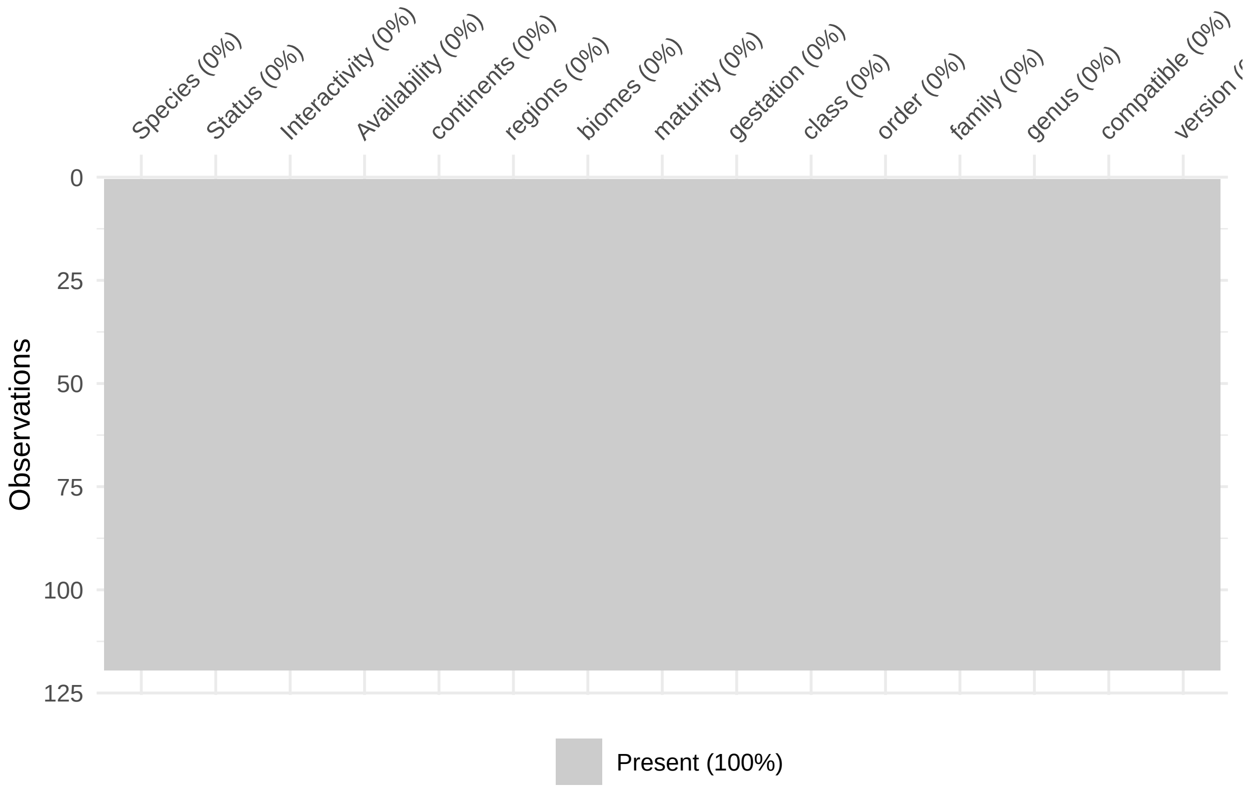
```
var_titles <-c("intercativity","version","continents","regions","biomes","maturity","gestation","class"
animal_table <- current_session %>% html_node('#mw-content-text') %>% html_node('table') %>% html_table
animal_table <- animal_table %>% mutate(continents=NA,
                                        regions = NA,
                                        biomes = NA,
                                        maturity = NA,
                                        gestation = NA,
                                        class = NA,
                                        order = NA,
                                        family = NA,
                                        genus = NA,
                                        image_link = NA,
                                        compatible = NA)
```

visit each individual animal page and scrape the required data note: to not overload the server with scrape requests a mandatory wait period of 2 seconds is used between requests

```
for (i in 1:nrow(animal_table)) {
  animal_name <- str_replace_all(animal_table$Species[i]," ","_")

  current_session <-current_session %>%
    session_jump_to (paste0("https://planetzoo.fandom.com/wiki/",animal_name))

  infobox_data <- current_session %>%
    html_node('aside')

  animal_image <- infobox_data %>%
    html_node('figure') %>%
    html_node('a') %>%
    html_attr('href')

  #download.file(animal_image,mode='wb',destfile = paste0("zooquiz/",animal_name,'.jpg'))
  animal_table[i,"version"] <- infobox_data %>%
    html_nodes(xpath='section[1]/div[2]/div') %>%
    html_text2() %>%
    paste(collapse =",")

  animal_table[i,"continents"] <- infobox_data %>%
    html_nodes(xpath='section[2]/div[1]/div') %>%
    html_text2() %>%
    paste(collapse =",")

  animal_table[i,"regions"] <- infobox_data %>%
    html_nodes(xpath='section[2]/div[2]/div') %>%
    html_text2()

  animal_table[i,"Status"] <- infobox_data %>%
    html_nodes(xpath='section[2]/div[3]/div/a/img') %>%
    html_attr("alt") %>%
    paste(collapse =",")

  animal_table[i,"biomes"] <- infobox_data %>%
    html_nodes(xpath='section[3]/section[4]/section[2]/div/a') %>%
    html_attr("title") %>%
    paste(collapse =",")

  animal_table[i,"maturity"] <- infobox_data %>%
    html_nodes(xpath='section[5]/section[2]/section[2]/div[1]') %>%
    html_text2() %>%
    paste(collapse =",")

  animal_table[i,"gestation"] <- infobox_data %>%
    html_nodes(xpath='section[5]/section[3]/section[2]/div[1]') %>%
    html_text2() %>%
    paste(collapse =",")

  animal_table[i,"class"] <- infobox_data %>%
    html_nodes(xpath='section[6]/section[1]/section[2]/div[1]') %>%
    html_text2() %>%
    paste(collapse =",")
  animal_table[i,"order"] <- infobox_data %>%
    html_nodes(xpath='section[6]/section[1]/section[2]/div[2]') %>%
    html_text2() %>%
    paste(collapse =",")

  animal_table[i,"family"] <- infobox_data %>%
    html_nodes(xpath='section[6]/section[2]/section[2]/div[1]') %>%
    html_text2() %>%
    paste(collapse =",")

  animal_table[i,"genus"] <- infobox_data %>%
    html_nodes(xpath='section[6]/section[2]/section[2]/div[2]') %>%
    html_text2() %>%
    paste(collapse =",")

  if (animal_table[i,"Interactivity"]=="Full"){
    animal_table[i,"compatible"] <- current_session %>%
      html_node('.compatible') %>%
      html_node('.compatible-text') %>%
      html_nodes('a') %>%
      html_attr('title') %>%
      paste(collapse =",")
  }
  #animal_table[i,"image_link"] <- paste0("zooquiz/",animal_name,'.jpg')
  Sys.sleep(2)

}
```

using the naniar library, we can quickly assess the amount of missing values in the data (and thus the quality of the scrape)

```
library(naniar)
vis_miss(animal_table%>%select(!image_link)%>%filter(Interactivity=='Full'))
```



glimpse the data to display the structure of the resulting dataset

```
glimpse(animal_table)
```

```
## Rows: 158
## Columns: 16
## $ Species       <chr> "African Savannah Elephant", "Grizzly Bear", "West Afric~
## $ Status        <chr> "Endangered", "Least concern", "Critically endangered", ~
## $ Interactivity <chr> "Full", "Full", "Full", "Full", "Full", "Full", "Full", ~
## $ Availability  <chr> "Standard", "Standard", "Standard", "Standard", "Standar~
## $ continents    <chr> "Africa", "North America", "Africa", "Africa", "Africa",~
## $ regions       <chr> "Sub Saharan Africa: Kenya, Tanzania, Botswana, Zimbabwe~
## $ biomes        <chr> "Desert,Grassland", "Taiga,Temperate,Tundra", "Grassland~
## $ maturity      <chr> "15 years", "8 years", "3 years", "4 years", "6 years", ~
## $ gestation     <chr> "22 months", "8 months", "3 months", "13 months", "8 mon~
## $ class         <chr> "Mammalia", "Mammalia", "Mammalia", "Mammalia", "Mammali~
## $ order         <chr> "Proboscidea", "Carnivora", "Carnivora", "Perissodactyla~
## $ family        <chr> "Elephantidae", "Ursidae", "Felidae", "Equidae", "Hippop~
## $ genus         <chr> "Loxodonta", "Ursus", "Panthera", "Equus", "Hippopotamus~
## $ image_link    <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ compatible    <chr> "", "", "", "African Buffalo,Black Wildebeest,Blue Wilde~
## $ version       <chr> "Standard", "Standard", "Standard", "Standard", "Standar~
```

save the completed data file into CSV format for use in future appliactions and files

```
write.csv(animal_table,file = "ZooQuiz.csv")
```