Group project-part 1

Emily, David, Daisy, and Jingxin

2024-11-01

Summary:

Interests:

We are a group of students passionate about environmental issues. While exploring datasets from the United Nations, the "Environment" section caught our eye, specifically the data on threatened species. Understanding species loss can offer valuable insights into environmental history, reveal ongoing extinction patterns, and identify specific groups (vertebrates, invertebrates, or plants) that may require more focused conservation efforts. Since ecosystems are highly interconnected, examining trends in species loss may also help us predict future impacts and highlight urgent actions needed to prevent further biodiversity decline.

Data Set:

The threatened species dataset from the UN Environment section.

Question:

In our initial dataset analysis, we noticed a correlation between the number of species threatened and changes over time. For this project, we aim to explore this further by investigating the question:

"How has the number of threatened species changed over time?"

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(corrplot)
```

```
## corrplot 0.94 loaded
```

Part 1:

- data info
- how many unknown data there is
- clean unknown data

```r
# The first row is a header so we want to skip it
df <- read_csv("endangered_species.csv", skip = 1)
```

```
## New names:
## Rows: 6921 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (4): ...2, Series, Footnotes, Source dbl (2): Region/Country/Area, Year num
## (1): Value
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...2`
```

```r
# There's 6921 rows and 7 columns
head(df)
```

```
## # A tibble: 6 x 7
##   `Region/Country/Area` ...2         Year Series          Value Footnotes Source
##                   <dbl> <chr>       <dbl> <chr>           <dbl> <chr>     <chr>
## 1                     4 Afghanistan  2004 Threatened Spe~    31 <NA>      World~
## 2                     4 Afghanistan  2010 Threatened Spe~    31 <NA>      World~
## 3                     4 Afghanistan  2015 Threatened Spe~    31 <NA>      World~
## 4                     4 Afghanistan  2019 Threatened Spe~    33 <NA>      World~
## 5                     4 Afghanistan  2020 Threatened Spe~    33 <NA>      World~
## 6                     4 Afghanistan  2021 Threatened Spe~    38 <NA>      World~
```

```r
glimpse(df)
```

```
## Rows: 6,921
## Columns: 7
## $ `Region/Country/Area` <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ~
## $ ...2                  <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Af~
## $ Year                  <dbl> 2004, 2010, 2015, 2019, 2020, 2021, 2022, 2004, ~
## $ Series                <chr> "Threatened Species: Vertebrates (number)", "Thr~
## $ Value                 <dbl> 31, 31, 31, 33, 33, 38, 42, 1, 1, 2, 2, 2, 2, 2,~
## $ Footnotes             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Source                <chr> "World Conservation Union (IUCN), Gland and Camb~
```

- So all the variables are locations, year, series (in particular, what type of species it is, vertebrates, invertebrates, or plants), then there's numerical values, footnote, and source of information

```r
# How many NA value there are
colSums(is.na(df))
```

```
## Region/Country/Area                ...2                Year              Series
##                   0                   0                   0                   0
##               Value           Footnotes              Source
##                   0                6807                   0
```

```r
# What's the percentage
sapply(df, function(x) mean(is.na(x)) * 100)
```

```
## Region/Country/Area                ...2                Year              Series
##             0.00000             0.00000             0.00000             0.00000
##               Value           Footnotes              Source
##             0.00000            98.35284             0.00000
```

```r
# Drop Footnotes
df <- df %>% select(-Footnotes)
```

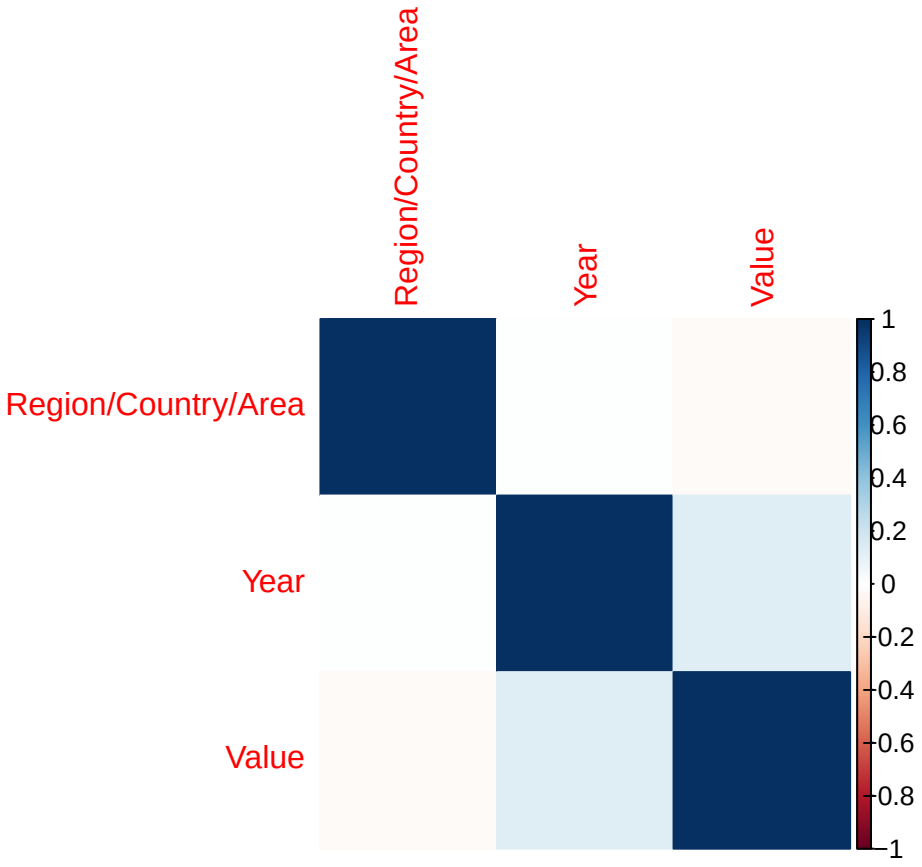
Correlation graph: - This is just a draft graph - trying to see a basic relationship of the variables

```r
# select numerical data
numeric_df <- df %>% select_if(is.numeric)

# calculate the matrix
cor_matrix <- cor(numeric_df, use = "complete.obs")

# adjust the margins
par(mar = c(1, 1, 2, 5))

# plot with
corrplot(cor_matrix, method = "color")
```



What we see is that there seems to be a correlation between "Year" and "Value".