Midterm1

Hector Carvajal

2025-02-07

In this midterm we will analyze some data on the conservation status of species in North America and spending under the Endangered Species Act.

Answer the following questions by using chunks of R code. Comment on what your code does. Make sure to add informative axis titles and, where appropriate, units to your answers. Upload the R markdown file and knitted output to Canvas.

We will use the file `conservationdata.csv`. This dataset has information on North American species. It has five variables that are described in the table below.

Table 1: Table 1. Variables in "consevationdata.csv"

Name	Description
speciesid	unique ID
speciesname	scientific name
taxon	Species group
conservation status	Conservation status in North America, according to NatureServe: 1 = Critically Imperiled; 2 = Imperiled; 3 = Vulnerable; 4 = Apparently Secure; 5 = Secure; UNK = Unknown; Prob. Extinct = Probably Extinct; Extinct
listed	Is the species listed as threatened or endangered under the US Endangered Species Act: 0 = No; 1 = Yes

Read in the file `conservationdata.csv`

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#reading in the data
conservation_data <- read_csv("conservationdata.csv")
```

```
## Rows: 53658 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): speciesname, taxon, conservation_status
## dbl (2): speciesid, listed
##
## 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.
```

```
#views the first few rows
head(conservation_data)
```

```
## # A tibble: 6 x 5
##   speciesid speciesname           taxon         conservation_status listed
##       <dbl> <chr>                 <chr>         <chr>                <dbl>
## 1     40733 Centroptilum sp. 1    Invertebrates UNK                      0
## 2     47278 Crosbyella spinturnix Invertebrates UNK                      0
## 3     47272 Lirceus garmani       Invertebrates UNK                      0
## 4     47273 Lirceus hoppinae      Invertebrates UNK                      0
## 5     47274 Lirceus ouachitaensis Invertebrates UNK                      0
## 6     47275 Lirceus trilobus      Invertebrates UNK                      0
```

```
conservation_data
```

```
## # A tibble: 53,658 x 5
##    speciesid speciesname           taxon         conservation_status listed
##        <dbl> <chr>                 <chr>         <chr>                <dbl>
##  1     40733 Centroptilum sp. 1    Invertebrates UNK                      0
##  2     47278 Crosbyella spinturnix Invertebrates UNK                      0
##  3     47272 Lirceus garmani       Invertebrates UNK                      0
##  4     47273 Lirceus hoppinae      Invertebrates UNK                      0
##  5     47274 Lirceus ouachitaensis Invertebrates UNK                      0
##  6     47275 Lirceus trilobus      Invertebrates UNK                      0
##  7     47277 Loxosceles reclusa    Invertebrates UNK                      0
##  8     47276 Macrocera nobilis     Invertebrates UNK                      0
##  9     47279 Trigenotyla blacki    Invertebrates 1                        0
## 10     47281 Caecidotea montana    Invertebrates UNK                      0
## # i 53,648 more rows
```

```
summary(conservation_data)
```

```
##    speciesid     speciesname           taxon           conservation_status
##  Min.   :    1   Length:53658       Length:53658       Length:53658      
##  1st Qu.:13415   Class :character   Class :character   Class :character  
##  Median :26830   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :26830                                                           
##  3rd Qu.:40244                                                           
##  Max.   :53658                                                           
##      listed       
##  Min.   :0.00000  
##  1st Qu.:0.00000  
##  Median :0.00000  
##  Mean   :0.03014
##  3rd Qu.:0.00000
##  Max.   :1.00000
```

1. What fraction of species in the dataset are listed under the Endangered Species Act? (2 points)

```
#count the species that are listed
sum(conservation_data$listed == 1)
```

```
## [1] 1617
```

```
#count total number of species
nrow(conservation_data)
```

```
## [1] 53658
```

```
#calculates the fraction of the species that are listed.
fraction_listed <- sum(conservation_data$listed == 1) / nrow(conservation_data)
print(fraction_listed)
```

```
## [1] 0.0301353
```

2. Show how many (absolute and relative) species there are for each taxonomic group by making a data.frame in which the first column has the name of the taxonomic groups, the second column is the number of species in that group, and the third column is the number of species in that group as a fraction of the total number of species in the dataset.

```
library(dplyr)
```

```
#calculate absolute and relative frequencies for each taxonomic group
taxon_summary <- conservation_data %>%
  group_by(taxon) %>%
  summarize(
#counts the number of species in each group
    species_count = n(),
#fraction of total speces
    fraction_total = species_count / nrow(conservation_data)
  ) %>%
  arrange(desc(species_count))
taxon_summary
```

```
## # A tibble: 9 x 3
##   taxon         species_count fraction_total
##   <chr>                 <int>          <dbl>
## 1 Invertebrates         24407        0.455
## 2 Plants                19511        0.364
## 3 Fungi                  6270        0.117
## 4 Fishes                 1453        0.0271
## 5 Birds                   795        0.0148
## 6 Mammals                 474        0.00883
## 7 Reptiles                350        0.00652
## 8 Amphibians              319        0.00595
## 9 Protists                 79        0.00147
```

```r
#turn it into a dataframe
taxon_summary_df <- as.data.frame(taxon_summary)
taxon_summary_df
```

```
##           taxon species_count fraction_total
## 1 Invertebrates         24407    0.454862276
## 2        Plants         19511    0.363617727
## 3         Fungi          6270    0.116851169
## 4        Fishes          1453    0.027078907
## 5         Birds           795    0.014816057
## 6       Mammals           474    0.008833725
## 7      Reptiles           350    0.006522793
## 8    Amphibians           319    0.005945059
## 9      Protists            79    0.001472287
```

3a) One interesting question is how the conservation status varies between different taxonomic groups. Make a plot showing the relative distribution of conservation status within each taxonomic group. There should be descriptive legend (with words, not with the numeric codes) (3 points)

You can use a "base" plotting method, or ggplot.

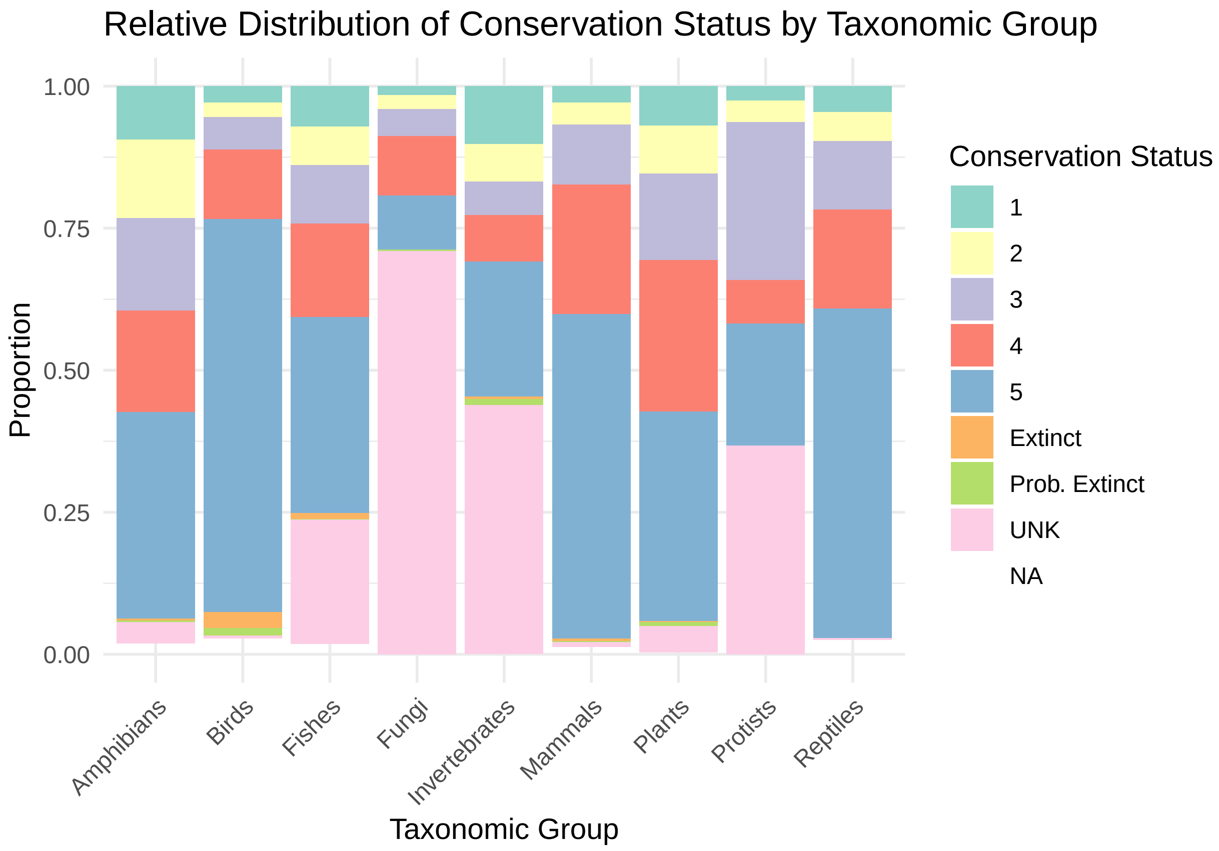
```r
library(dplyr)
library(ggplot2)

#counts species by taxon and conservation status
conservation_summary <- conservation_data %>%
  group_by(taxon, conservation_status) %>%
  summarize(count = n()) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'taxon'. You can override using the
## `.groups` argument.
```

```r
#converts to relative frequency
conservation_summary <- conservation_summary %>%
  group_by(taxon) %>%
  mutate(fraction = count / sum(count))
conservation_summary
```

```
## # A tibble: 74 x 4
## # Groups:   taxon [9]
##    taxon      conservation_status count fraction
##    <chr>      <chr>               <int>    <dbl>
##  1 Amphibians 1                      30  0.0940
##  2 Amphibians 2                      44  0.138
##  3 Amphibians 3                      52  0.163
##  4 Amphibians 4                      57  0.179
##  5 Amphibians 5                     116  0.364
##  6 Amphibians Extinct                 1  0.00313
##  7 Amphibians Prob. Extinct           1  0.00313
##  8 Amphibians UNK                    12  0.0376
##  9 Amphibians <NA>                    6  0.0188
## 10 Birds      1                      23  0.0289
## # i 64 more rows
```

```r
ggplot(conservation_data, aes(x = taxon, fill = conservation_status)) +
  geom_bar(stat = "count", position = "fill") +  #counts species per status
  labs(
    title = "Relative Distribution of Conservation Status by Taxonomic Group",
    x = "Taxonomic Group",
    y = "Proportion",
    fill = "Conservation Status"
  ) +
#color scheme
  scale_fill_brewer(palette = "Set3") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



3b) Based on this graph, what is something we might be concerned about in terms of analyzing the data on conservation status, particularly for fungi and invertebrates? (1 point)

Answer: A concerning factor to this is that we don't have enough data on fungi and invertebrates given that they have an "Unknown conservation status." Some of these species could be at risk but that is why this is concerning since we would not know.These group could have not been studied enough which would expain the lack of data.

Read in the second data file: `spendingdata.csv`

```r
library(readr)
spending_data <- read_csv("spendingdata.csv")
```

```
## Rows: 27630 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (3): speciesid, Year, spending
##
## 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.
```

```
head(spending_data)
```

```
## # A tibble: 6 x 3
##   speciesid  Year spending
##       <dbl> <dbl>    <dbl>
## 1     49476  2015  461813.
## 2     49476  2016  615705.
## 3     49477  2015  422095.
## 4     49477  2016  471121.
## 5        11  2014  956187.
## 6        11  2015  917125.
```

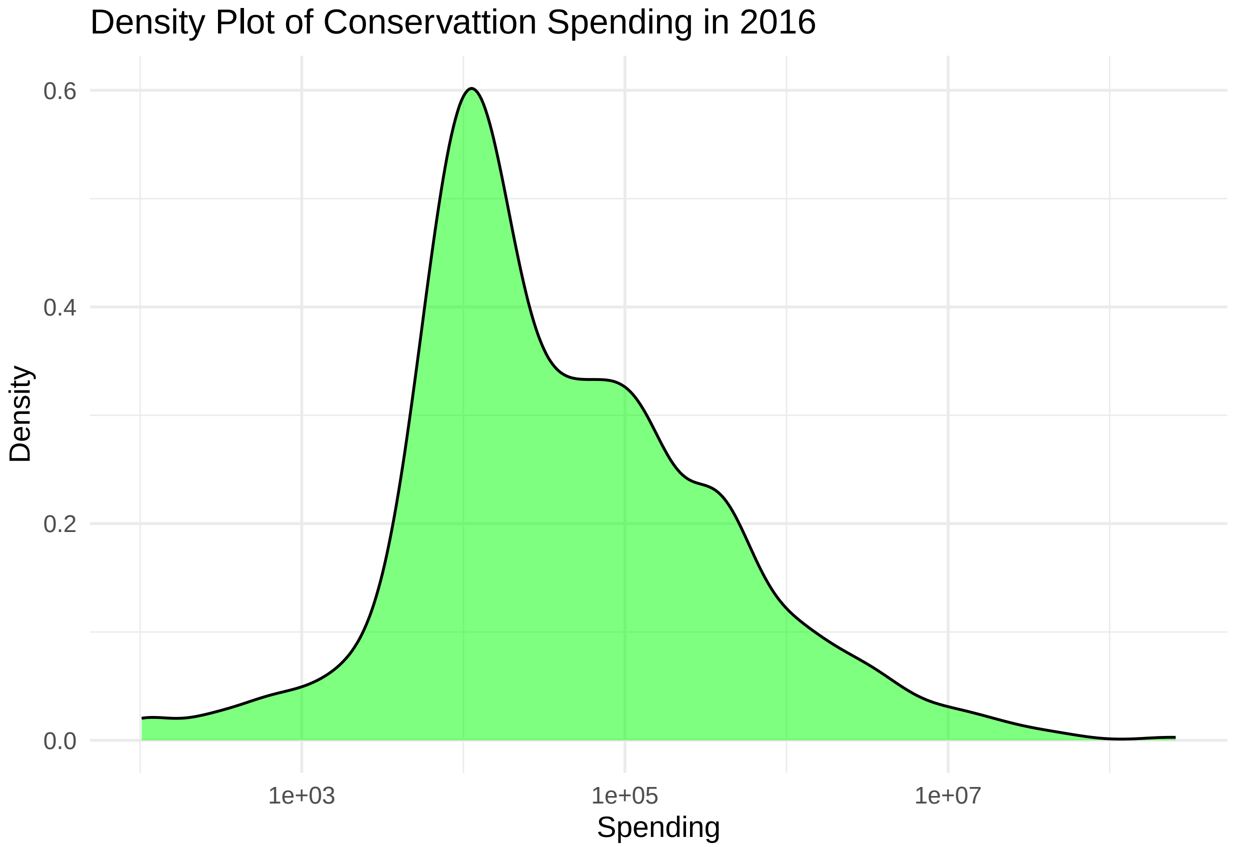
This dataset has a species ID that matches the species ID in the conservation dataset (speciesid), year, and the spending on conservation of that species (expressed in in 2015 dollars, i.e., accounting for inflation)

4a) Make a plot showing the distribution of spending in the year 2016 (3 points)

```
#filters spending data for 2016
spending_2016 <- spending_data %>%
  filter(Year == 2016)
head(spending_2016)
```

```
## # A tibble: 6 x 3
##   speciesid  Year spending
##       <dbl> <dbl>    <dbl>
## 1     49476  2016  615705.
## 2     49477  2016  471121.
## 3        11  2016 1014963.
## 4        54  2016 1073824.
## 5        84  2016    1838.
## 6        94  2016  913666.
```

```
ggplot(spending_2016, aes(x = spending)) +
  geom_density(fill = "green", alpha = 0.5) +
  scale_x_log10() +  #log here helps with dealing with skewed data
  labs(
    title = "Density Plot of Conservattion Spending in 2016",
    x = "Spending",
    y = "Density"
  ) +
  theme_minimal()
```



4b) Notice the (very) long right tail on spending data - we spend a lot on a very small number of species. Show the IDs of the 3 species with the most spending in 2016. (2 points)

```
#top 3 species that received the most conservation funding in 2016
top_species_2016 <- spending_2016 %>%
#soreted bt highest spending
  arrange(desc(spending)) %>%
#simple way of showing the first three
   head(3)
top_species_2016
```

```
## # A tibble: 3 x 3
##   speciesid  Year   spending
##       <dbl> <dbl>      <dbl>
## 1      1632  2016 255893066.
## 2      4486  2016 229175092.
## 3      1684  2016  54122671.
```

5. Merge in the data from the conservation status data frame to the spending data frame, so that we have information on species names, taxonomic group, and conservation status with the spending data. (2 points); and use that to show the scientific names of the three species identified above.

```
#merged data with the spending data
merged_data <- spending_data %>%
  left_join(conservation_data,by = "speciesid")
head(merged_data)
```

```
## # A tibble: 6 x 7
##   speciesid  Year spending speciesname         taxon  conservation_status listed
##       <dbl> <dbl>    <dbl> <chr>               <chr>  <chr>                <dbl>
## 1     49476  2015  461813. Orbicella faveolata Inver~ 2                        1
## 2     49476  2016  615705. Orbicella faveolata Inver~ 2                        1
## 3     49477  2015  422095. Orbicella franksi   Inver~ 3                        1
## 4     49477  2016  471121. Orbicella franksi   Inver~ 3                        1
## 5        11  2014  956187. Balaena mysticetus  Mamma~ 3                        1
## 6        11  2015  917125. Balaena mysticetus  Mamma~ 3                        1
```

```
#scientific names of top 3 species identified
top_species_withNames <- merged_data %>%
  filter(speciesid %in% top_species_2016$speciesid) %>%
  select(speciesid, speciesname, taxon, conservation_status, spending)  # Select useful columns
top_species_withNames
```

```
## # A tibble: 60 x 5
##    speciesid speciesname              taxon  conservation_status   spending
##        <dbl> <chr>                    <chr>  <chr>                    <dbl>
##  1      1632 Oncorhynchus tshawytscha Fishes 5                   230821991.
##  2      1632 Oncorhynchus tshawytscha Fishes 5                   281448714.
##  3      1632 Oncorhynchus tshawytscha Fishes 5                   255893066.
##  4      1632 Oncorhynchus tshawytscha Fishes 5                   124462342.
##  5      1632 Oncorhynchus tshawytscha Fishes 5                    88365223.
##  6      1632 Oncorhynchus tshawytscha Fishes 5                    88560494.
##  7      1632 Oncorhynchus tshawytscha Fishes 5                    90394631.
##  8      1632 Oncorhynchus tshawytscha Fishes 5                   129321364.
##  9      1632 Oncorhynchus tshawytscha Fishes 5                   160693129.
## 10      1632 Oncorhynchus tshawytscha Fishes 5                   185133533.
## # i 50 more rows
```

Look up these scientific names - what is the common name for these species?

Answer: The names that result after searching theses species name up is Oncorhynchus tshawytscha which has a common name: Chinook Salmon, Oncorhynchus kisutch which has the common name: Coho Salmon, and finally Oncorhynchus mykiss which has the common name: Rainbow Trout.

6. Finally, we will use a regression to look at the relationship between spending and species taxon.

Because the distribution of spending is very right-skewed, it would be a good idea to take the logarithm of spending before using it in a regression.

Remember that log(0)=infinity. That means we have to drop observations with zero spending before taking the logarithm.

a) Drop the rows where spending == 0 from the data frame and then make a new column with the logarithm (log()) of spending in each year. (2 points)

```
#removing rows where spending is 0
cleaned_data <- merged_data %>%
  filter(spending > 0) %>%
  mutate(log_spending = log(spending))
summary(cleaned_data$log_spending)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.987   9.146  10.422  10.638  12.046  19.562
```

Optional: Look at the distribution of the logged spending variable and see how it looks different from the plot you made in question 4a

b) Run a regression of logged spending on taxonomic group and print the summary for the regression below (3 points)

```
#regression of log-transformed spending on taxon
spending_model <- lm(log_spending ~ taxon, data = cleaned_data)
summary(spending_model)
```

```
## 
## Call:
## lm(formula = log_spending ~ taxon, data = cleaned_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.7311 -1.1848  0.0171  1.3813  7.4867 
## 
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          11.64222    0.09488 122.700  < 2e-16 ***
## taxonBirds            0.87617    0.10555   8.301  < 2e-16 ***
## taxonFishes           0.43339    0.10266   4.222 2.43e-05 ***
## taxonFungi           -1.63702    0.32276  -5.072 3.97e-07 ***
## taxonInvertebrates -0.64918    0.09927  -6.540 6.28e-11 ***
## taxonMammals          1.03077    0.10690   9.643  < 2e-16 ***
## taxonPlants          -1.92320    0.09628 -19.975  < 2e-16 ***
## taxonReptiles         0.48029    0.12093   3.972 7.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.999 on 26963 degrees of freedom
## Multiple R-squared:  0.2402, Adjusted R-squared:   0.24 
## F-statistic:  1218 on 7 and 26963 DF,  p-value: < 2.2e-16
```

c) The way to interpret these coefficients are as the fractional difference in spending between the taxonomic group (e.g. Birds, Fishes etc) and the "dropped" group, where by default the dropped group will be Amphibians. Positive numbers indicate that group has more spent on it than Amphibians and negative numbers indicate it has less spent on it.

Based on your results in b, do we see statistically significant differences in spending between different taxonomic groups? If so, which kinds of species tend to have more spent on them and which have less? (1 points)

Answer: The output shown in part b reveals that the regression results show significant difference in conservation's spending between taxonomic groups compared to amphibians. The ones receiving most of the fundung are mammals, birds, fishes and reptiles which have positive coefficients. The lower funded groups are fungi, invertebrates and plants. Mammals revieves the most while plants receive the least.

7. Push your R markdown file to your Github repository (2 points)