Midterm One

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`

```
#Reading the conservation data into R
conservation <- read.csv("conservationdata.csv", stringsAsFactors = FALSE)
```

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

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

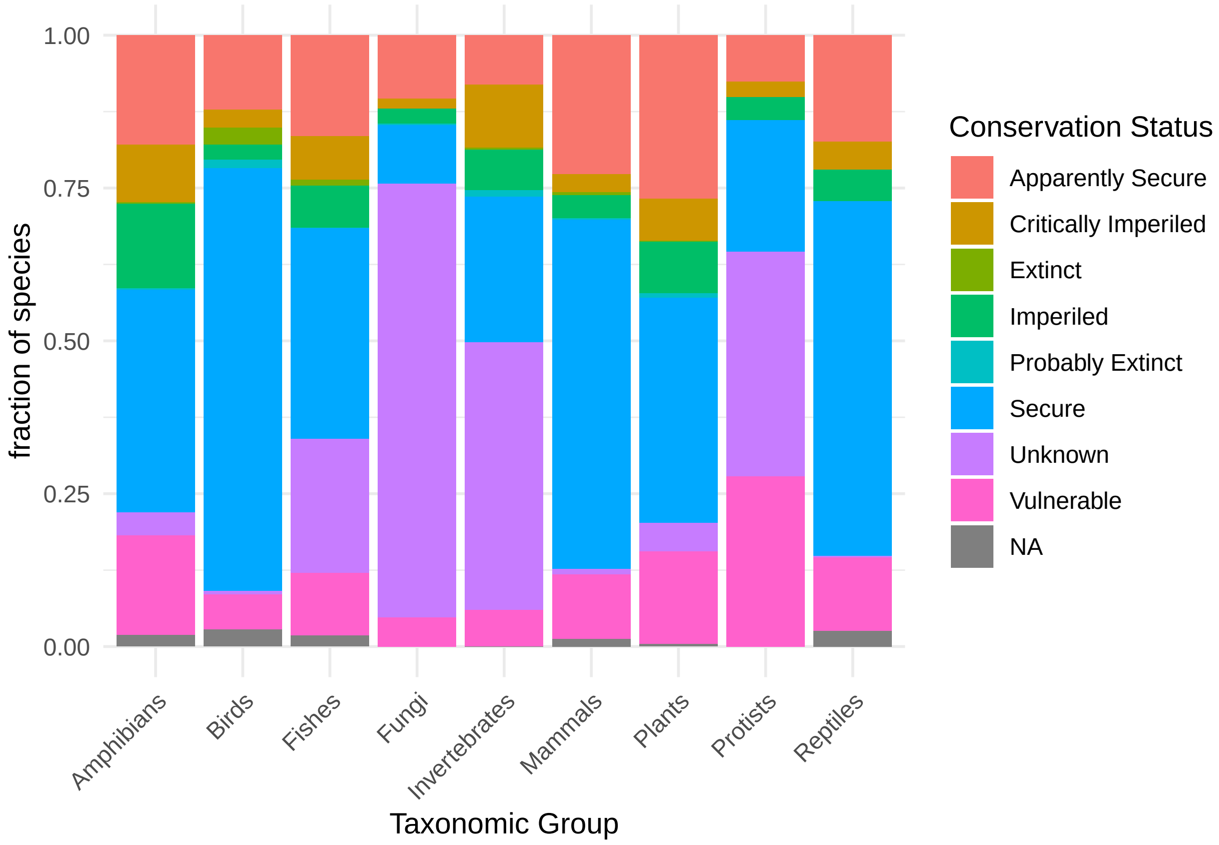
2. Show how many (absolute and relative) species there are for each taxonomic group by making a data.frame in which the first column has the name of the taxonomic groups, the second column is the number of species in that group, and the third column is the number of species in that group as a fraction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

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

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.

If you are using ggplot, stat="count" (counts up and plots the number of observations, i.e. species, within each group) and position="fill" might both be useful.



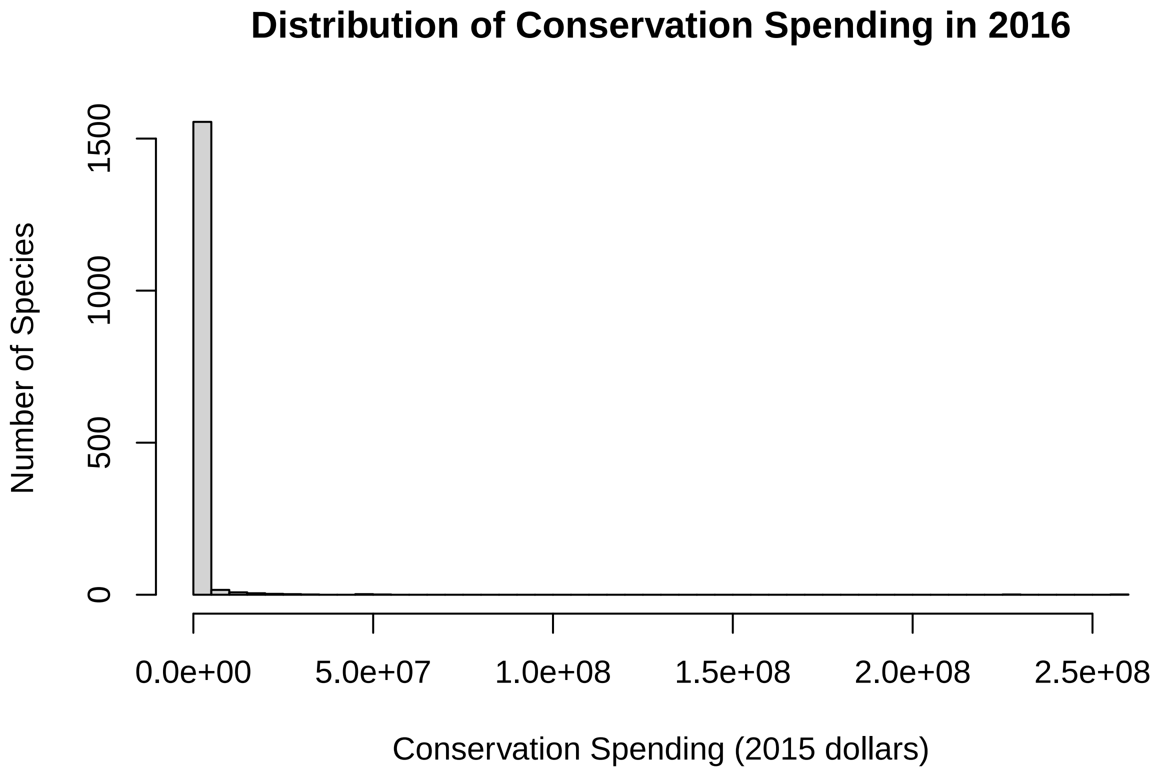
3b) Based on this graph, what is something we might be concerned about in terms of analyzing the data on conservation status, particularly for fungi and invertebrates? (1 point)

Answer: For fungi and invertebrates, we should be concerned about the high percentage of 'unknown' within those particular taxonomies. In regards to conservation status, analysis would be difficult given the huge unknown status, which may lead to improper (and even lack of) conservation efforts.

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

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)



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)

```
## [1] 1632 4486 1684
```

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.

```
##      speciesid Year  spending              speciesname  taxon
## 2191      1632 2016 255893066 Oncorhynchus tshawytscha Fishes
## 2316      1684 2016  54122671     Oncorhynchus kisutch Fishes
## 4744      4486 2016 229175092      Oncorhynchus mykiss Fishes
##      conservation_status listed status_label
## 2191                   5      1       Secure
## 2316                   5      1       Secure
## 4744                   5      1       Secure
```

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

Answer: Chinook Salmon, Coho Salmon, and 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)

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)

```
##
## Call:
## lm(formula = log_spending ~ taxon, data = spending_nonzero)
##
## 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: Yes, there is an obvious statistical and significant difference in spending between different taxonomic groups. For instance, when compared to amphibians, birds and mammals tended to received much more of a substantial fund for conservation with mammals also having the largest positive difference (and then birds). Fishes and reptiles also received significantly more funding than amphibians, though to a lesser extent. However, when it comes to plants, fungi, and invertebrates, these taxonomic groups received far less conservation spending with plants receiving the lowest relative funding.

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