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`

```
conservation_data <- read.csv("conservationdata.csv")
```

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

```
endangered_list <- table(conservation_data$listed)
##used table function to find number of zeros in listed column

total_fraction_listed <- 1617/53658
##divided number of ones in column by total number of species

endangered_list_alt <- mean(conservation_data$listed)
##alternatively used mean function to find fraction because listed column uses 0's and 1's so when adde
endangered_list_alt
```

```
## [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.

```
##find absolute number
counts <- table (conservation_data$taxon)
##use data.frame find taxon groups and then bumber in each group and then fractions
taxon_summary <- data.frame(
  taxon_group = names(counts),
  number_of_species = as.vector(counts),
  fraction_of_total = as.vector(counts) / nrow(conservation_data)
)
taxon_summary
```

```
##     taxon_group number_of_species fraction_of_total
## 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.

```
##use ggplot and geom bar to create a distribution of taxon conservation status
ggplot(conservation_data, aes(x=taxon,fill = conservation_status))+geom_bar(position = "fill", stat="co
    x = "Taxonomic Group",
    y = "Proportion",
    fill = "Conservation Status")
```

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: Something to consider would be that are a much larger number and variance among invertebrates and fungi that makes it extremely difficult to track each and every single branch off.

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

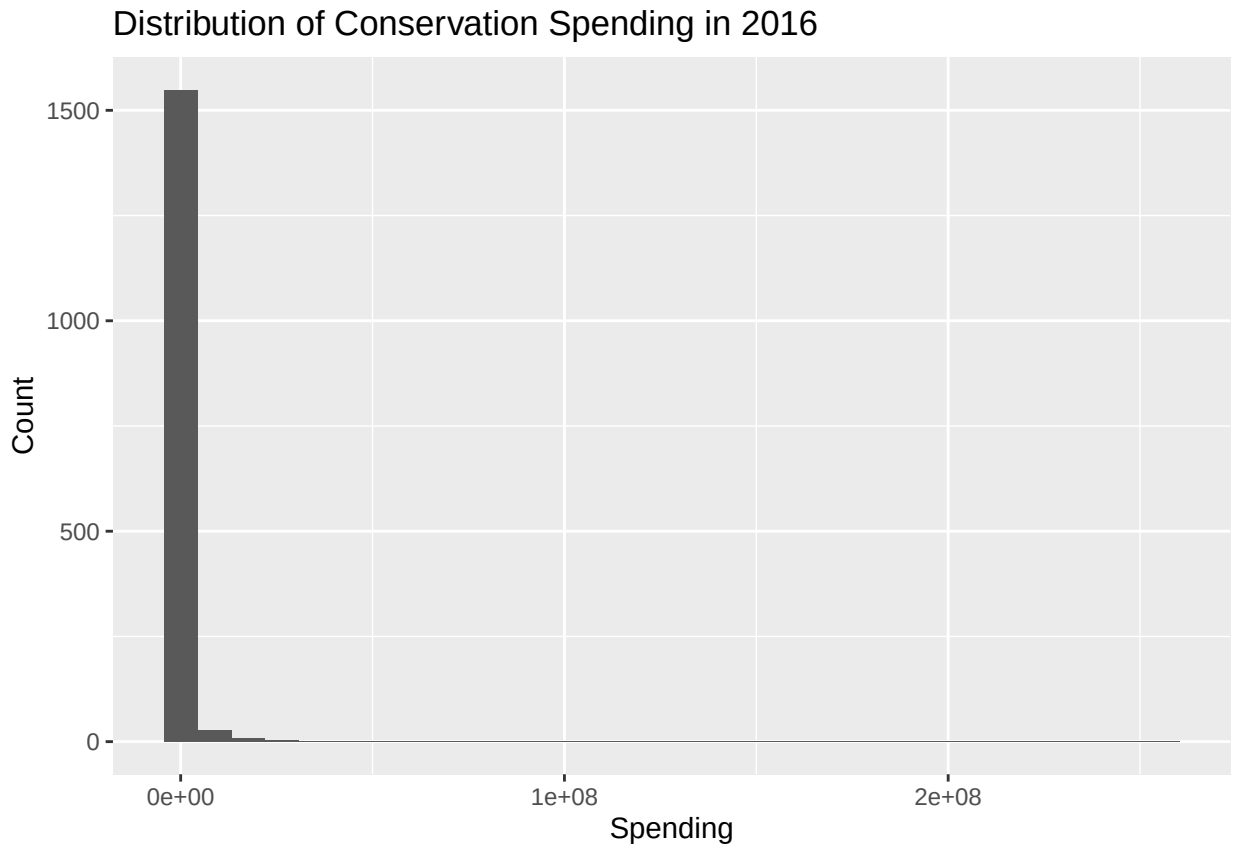
```
spending <- read.csv("spendingdata.csv")
```

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)

```
spending_2016 <- spending[spending$Year == 2016, ]
##make a data set with all the spending from only 2026
```

```r
ggplot(spending_2016, aes(x = spending)) +
  geom_histogram(bins = 30) +
  labs(
    title = "Distribution of Conservation Spending in 2016",
    x = "Spending",
    y = "Count"
  )
```




```r
##make a ggplot histogram to show a distrubutiion of data
```

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)

```r
##use decreasing = true to find the outliers
tops_3 <- spending_2016$speciesid[
  order(spending_2016$spending, decreasing = TRUE)[1:3]
]
```

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.

```r
#use common clumn name to merge
merged_data <- merge(
  spending,
  conservation_data,
  by = "speciesid"
)

##refind the 'top 3' from new merdged data set
top_3 <- subset(merged_data, Year == 2016)
top_3 <- top_3[order(-top_3$spending), ][1:3, ]

##print the species name of 'top 3'
top_3$speciesname
```

```
## [1] "Oncorhynchus tshawytscha" "Oncorhynchus mykiss"
## [3] "Oncorhynchus kisutch"
```

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

Answer: Oncorhynchus tshawytscha is known as King Salmon, Oncorhynchus mykiss is known as Rainbow Trout, and Oncorhynchus kisutch is also known as Silver Salmn.

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)

```r
## use subset to remove rows with spending =0
merged_data <- subset(merged_data, spending != 0)

##creat a new row withe th elogarithm of each spending year
merged_data$logspending <- log(merged_data$spending)
```

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)

```r
mod <- lm(logspending~taxon, data = merged_data)
sum_mod<-summary(mod)
sum_mod$coefficients
```

```
##                      Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)        11.6422242 0.09488366 122.699988 0.000000e+00
```

```
## taxonBirds           0.8761714 0.10554856   8.301121 1.079041e-16
## taxonFishes          0.4333917 0.10265748   4.221725 2.432332e-05
## taxonFungi          -1.6370214 0.32276388  -5.071885 3.965026e-07
## taxonInvertebrates -0.6491766 0.09926884  -6.539580 6.279492e-11
## taxonMammals         1.0307675 0.10689705   9.642619 5.731912e-22
## taxonPlants         -1.9231981 0.09628168 -19.974705 3.972023e-88
## taxonReptiles        0.4802943 0.12093378   3.971547 7.159261e-05
```

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 statistical difference shown between spending among all taxons- birds and fish receive the most spending.

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