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)

```
ncol(conservation_data)
```

```
## [1] 5
```

```
colnames(conservation_data)
```

```
## [1] "speciesid"           "speciesname"         "taxon"
## [4] "conservation_status" "listed"
```

```
ESAfraction=mean(conservation_data$listed,na.rm = 1)
#code takes the fraction of species that are listed under the Endangered Species Act, under listed colu
```

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.

```r
species_taxgroup=table(conservation_data$taxon)
tax_groups=as.data.frame(species_taxgroup)
colnames(tax_groups)=c("name","numberofspecies")
tax_groups$fraction=tax_groups$numberofspecies/sum(tax_groups$numberofspecies)
tax_groups
```

```
##            name numberofspecies    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
```

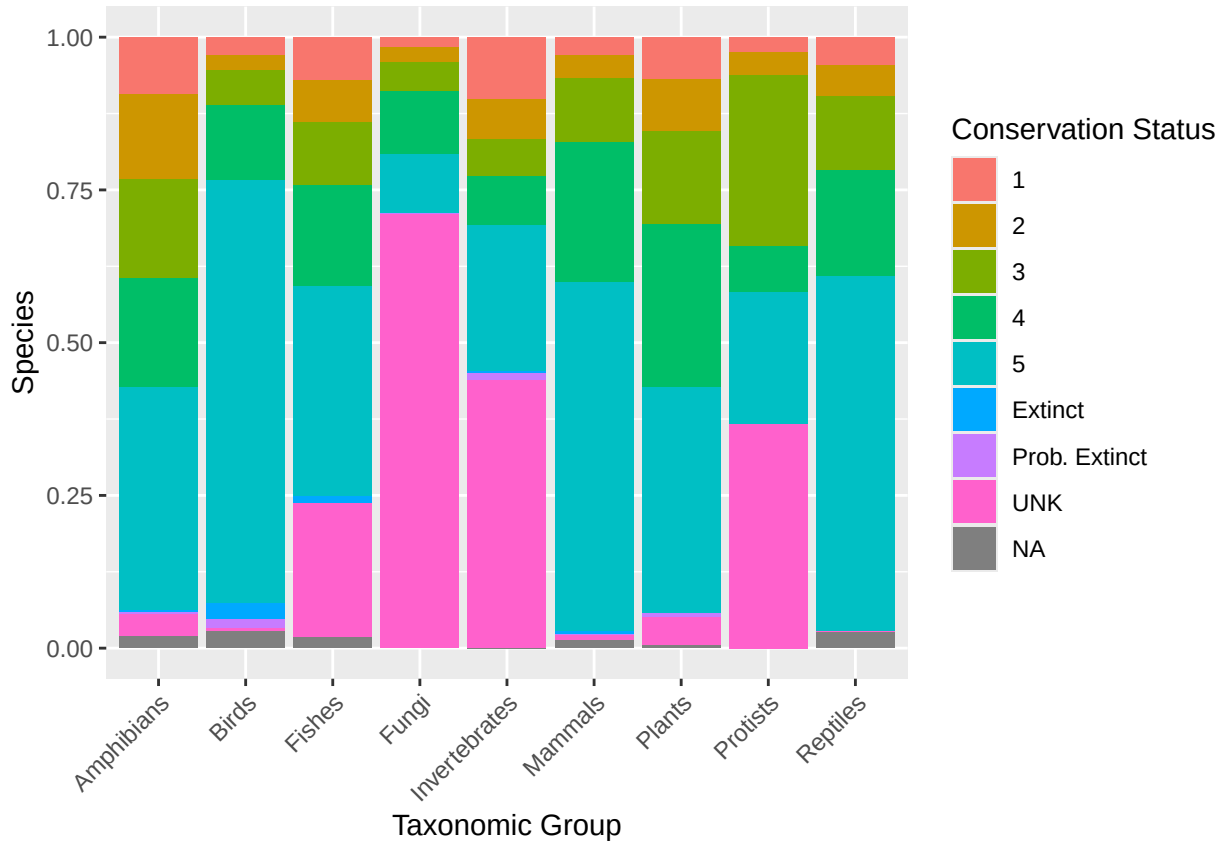
```r
#first code establishes table of number of species and their respective taxonomic groups. Next I create
```

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.

```r
library(ggplot2)
ggplot(conservation_data)+geom_bar(aes(x=.data$taxon,fill=.data$conservation_status),stat="count",positi
```




```
#here I am making a plot of the conservation_data using ggplot, first I set the x-axis to be the differ
```

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:The graph shows the major distribution of UNK values for both fungi and invertebrates. This highlights the huge gap in data for these taxonomic groups on their conservation status.

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)

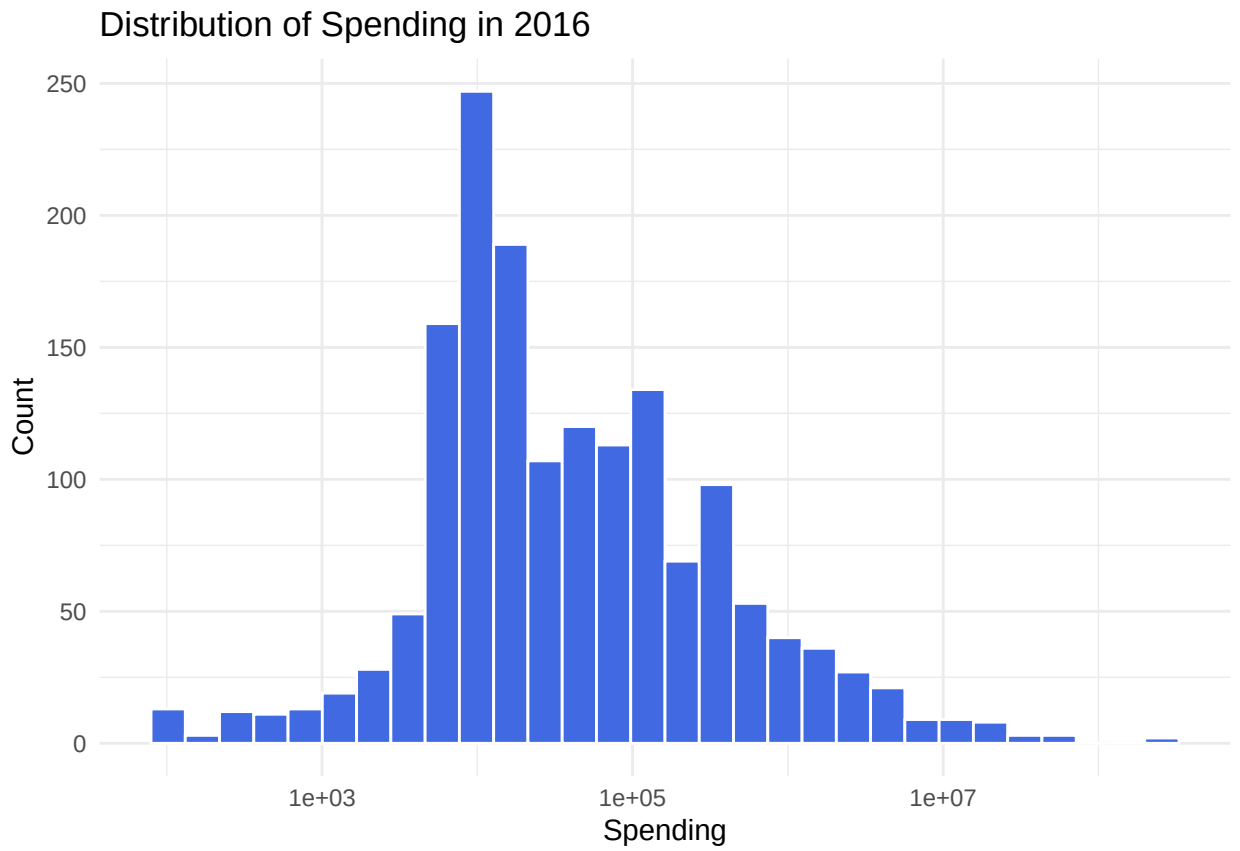
4a) Make a plot showing the distribution of spending in the year 2016 (3 points)

```
library(ggplot2)
spending_data=read.csv('spendingdata.csv')
names(spending_data)
```

```
## [1] "speciesid" "Year"      "spending"
```

```
spending_2016=spending_data[spending_data$Year==2016,]
ggplot(spending_2016, aes(x=spending))+
  geom_histogram(bins=30,fill="royalblue",color="white")+
  scale_x_log10()+
```

```
  labs(
    title="Distribution of Spending in 2016",
    x="Spending",
    y="Count"
  )+
theme_minimal()
```




```
str(spending_2016$spending)
```

```
##  num [1:1595] 615705 471121 1014963 1073824 1838 ...
```

```
spending_2016$spending=as.numeric(spending_2016$spending)
#First I created a new object that would hold only the 2016 spending data and called it spending_2016.
```

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=spending_2016[order(spending_2016$spending,decreasing=TRUE),]
top_3[1:3,c("speciesid","spending")]
```

```
##      speciesid  spending
## 2095      1632 255893066
## 4627      4486 229175092
## 2220      1684  54122671
```

```
#First i made top_3 to represent the spending_2016 data by decreasing spending data. Then i highlighted
```

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.

```
all_data=merge(
  spending_data,
  conservation_data,
  by="speciesid",
  all.x=TRUE
)
merge_2016=subset(all_data,Year==2016)
top3_2016=merge_2016[order(merge_2016$spending,decreasing=TRUE),][1:3, ]
top3_2016[,c("speciesid","speciesname")]
```

```
##      speciesid              speciesname
## 2191      1632 Oncorhynchus tshawytscha
## 4744      4486      Oncorhynchus mykiss
## 2316      1684     Oncorhynchus kisutch
```

```
#First i merged both the spending_data and conservation_data datasets as a whole, then created merge_20
```

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

Answer:The common name of speciesid:1632 is Chinook Salmon, the common name for speciesid 4486 is Rainbow trout, and thecommon name for species id 1684 is Coho Salmon

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)

```
regression=all_data[all_data$spending>0,]
regression$log_spending=log(regression$spending)
#Here I made a new dataset called regression of the all_data dataset without the 0 values for 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
regression2=lm(log_spending~taxon,data = regression)
summary(regression2)
```

```
##
## Call:
## lm(formula = log_spending ~ taxon, data = regression)
##
## 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:There are statistically significant differences in spending as some taxonomic groups have much more spending than others. Mammals and birds have by far the most conservation spending, meanwhile plants and fungi have the least sconservation spending by a wide margin.

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