Lab3

Matthew Valko

9/4/2019

Introduction

In this lab we return to the herbarium specimen data in which you will pick one Family and report on the following using what you learned in the last two weeks. Always handle missing data and justify why you chose the steps you did.

What is the distribution of years the specimen collected? Is it skewed? How do you know if you made the right choice? How did you handle missing data?

Supplemental I would like to know the median, IQR, mean and standard deviation of the year the specimen collected.

Is there a correlation between minimumElevationInMeters and startDayOfYear? Plot as well. Hint you might want to use na.omit to remove all missingness.

What are the top words associated the species you selected for the habitat field? Please show this graphically using a chart we discussed in lecture 2. Please clean the data for ’.’and stopwords using tidytext and stem the words using the SnowballC package. Please submit the lab report in pdf format. Does any of the words provide insight into the habitat of the Family you selected why or why not?

Load Packages

```
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
```

```
library(stringr)
library(SnowballC)
library(ggplot2)
library(tidytext)
```

Load data

```
df<-read.csv('~/Documents/CDS-303 Fall 2019/Lab/Lab3data.csv',header=T)
```

Year

Correlation

Word Processing

Word Graph