Lab2

Riley Payung

02/02/2020

Introduction

In this lab we will look at herbarium specimen data and use dplyr to perform data processing. Picture data is the URL link to the herbarium specimens. Occ data contains information on how the specimen was collected. We will be using “cuts” or sections of this data for labs for the rest of the semester. Herbarium specimens are collections of dried plant material overtime that aid botanists in the study of phenology or how species differe from each other. Always print your output for each section of your R markdown. Only user dplyr functions besides loading data and packages.

Load Data

```
occ = read.csv("occdata.csv")
pic = read.csv("picdata.csv")
```

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
```

Print the head of the data sets, and use the R base package to calculate the number of rows and columns in each data frame

	filenum <int>	coreid <int>	institutionCode <fctr>	collectionCode <fctr>	ownerInstitutionCode <lgl>
1	14	6763299	GMUF	Plants	*NA*

	filenum <int>	coreid <int>	institutionCode <fctr>	collectionCode <fctr>	ownerInstitutionCode <lgl>
2	14	6763300	GMUF	Plants	*NA*
3	14	6763352	GMUF	Plants	*NA*
4	14	6763375	GMUF	Plants	*NA*
5	14	6763419	GMUF	Plants	*NA*
6	14	6763478	GMUF	Plants	*NA*

6 rows | 1-6 of 83 columns

	id <int>	coreid <int>
1	15	7453510
2	15	7962893
3	15	7941303
4	15	6816246
5	15	8065784
6	15	7973098

6 rows | 1-3 of 20 columns

User an inner join between the two data sets using the commands in dplyr

	filenum <int>	coreid <int>	institutionCode <fctr>	collectionCode <fctr>	ownerInstitutionCode <lgl>
1	14	6763299	GMUF	Plants	*NA*
2	14	6763300	GMUF	Plants	*NA*
3	14	6763352	GMUF	Plants	*NA*
4	14	6763375	GMUF	Plants	*NA*
5	14	6763419	GMUF	Plants	*NA*
6	14	6763478	GMUF	Plants	*NA*

6 rows | 1-6 of 101 columns

Missing Data of Year of Collection(year), Count the number of missing data collection years.

```
## [1] 371
```

Based on our lecture and if I told you herbarium specimens have only been recorded for the last 100 years. Please fix the missingness of that column.

	filenum <int>	coreid <int>	institutionCode <fctr>	collectionCode <fctr>	ownerInstitutionCode <lgl>
1	14	6763299	GMUF	Plants	*NA*
2	14	6763300	GMUF	Plants	*NA*
3	14	6763352	GMUF	Plants	*NA*
4	14	6763375	GMUF	Plants	*NA*
5	14	6763419	GMUF	Plants	*NA*
6	14	6763478	GMUF	Plants	*NA*

6 rows | 1-6 of 101 columns

By phylum please take the average startDayOfYear which is when day of the year that it was collected

phylum <fctr>	avg <dbl>
Coniferophyta	160.3333
Lycopodiophyta	178.0714
Magnoliophyta	199.9498
Pteridophyta	221.0000

4 rows

Subset the phylum to Coniferophyta and print the dim of the data frame

```
## [1] 6
```

```
## [1] 100
```