Dirca\_Data\_Wrangling.R

Dirca

Fri Dec 14 12:19:34 2018

## ---------------------------------------------------------- ##  
# DIRCA ANATOMY DATA WRANGLING ####  
## ---------------------------------------------------------- ##  
  
##Install Libraries and Packages EX:install.packages("tidyverse")  
library(tidyverse) #loads ggplot2, tibble, tidyr, readr, purrr, dplyr, stringr, forcats

## ── Attaching packages ────────────────────────────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 3.0.0 ✔ purrr 0.2.5  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.6  
## ✔ tidyr 0.8.1 ✔ stringr 1.3.1  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ───────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(readxl) #hadley wickham's package for reading in excel files. not in Tidyverse.  
library(lubridate) #garett grolemund & hadley wickham's package for dates

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

setwd("/Users/Dirca/Desktop/GitHub-RStudio/Dirca Anatomy Mock Data Analysis/\_data/raw")  
#Load .csv dataset. Default: Header=TRUE  
#dirca <- read.csv("\_data/raw/dirca.csv", na.strings = c("", "NA", "na", " "))  
dirca <- read.csv("dirca.csv", na.strings = c("", "NA", "na", " "))  
  
#Explore dataset  
summary(dirca)

## species population specimen\_num cell\_type   
## decipiens :120 eureka : 60 Min. :1 bark\_fiber :96   
## mexicana :120 fl : 60 1st Qu.:1 companion\_cell :96   
## occidentalis: 60 isu : 60 Median :2 phloem\_parenchyma:96   
## palustris :180 jasper : 60 Mean :2 ray :96   
## me : 60 3rd Qu.:3 sieve\_tube\_mem :96   
## nd : 60 Max. :3   
## (Other):120   
## observation primary\_wall\_thickness secondary\_wall\_thickness  
## Min. :1.00 Min. : 12.00 Min. : 22.00   
## 1st Qu.:1.75 1st Qu.: 35.00 1st Qu.: 42.00   
## Median :2.50 Median : 48.00 Median : 53.00   
## Mean :2.50 Mean : 53.83 Mean : 54.75   
## 3rd Qu.:3.25 3rd Qu.: 65.00 3rd Qu.: 65.25   
## Max. :4.00 Max. :150.00 Max. :100.00   
## NA's :288   
## length lumen\_diameter cell\_total\_diameter lignified.  
## Min. : 70.0 Min. : 10.0 Min. : 45.0 no:480   
## 1st Qu.: 210.0 1st Qu.: 70.0 1st Qu.:150.0   
## Median : 270.0 Median :120.0 Median :176.5   
## Mean : 539.6 Mean :119.5 Mean :195.3   
## 3rd Qu.: 370.0 3rd Qu.:170.0 3rd Qu.:216.0   
## Max. :2900.0 Max. :240.0 Max. :400.0   
##   
## num\_of\_cell\_width seriate\_num   
## Min. :3.00 Min. :1   
## 1st Qu.:3.75 1st Qu.:1   
## Median :4.00 Median :1   
## Mean :4.00 Mean :1   
## 3rd Qu.:4.25 3rd Qu.:1   
## Max. :5.00 Max. :1   
## NA's :384 NA's :384

names(dirca)

## [1] "species" "population"   
## [3] "specimen\_num" "cell\_type"   
## [5] "observation" "primary\_wall\_thickness"   
## [7] "secondary\_wall\_thickness" "length"   
## [9] "lumen\_diameter" "cell\_total\_diameter"   
## [11] "lignified." "num\_of\_cell\_width"   
## [13] "seriate\_num"

###STEP 3###  
#Standardize/clean inconsistencies in column names, class etc.  
#Rename column headings  
#Use the rename function in the dplyr package  
#New name is on left, old name is on right.  
dirca <- rename(dirca, Species = species)  
dirca <- rename(dirca, Population = population)  
dirca <- rename(dirca, Specimen.Number = specimen\_num)  
dirca <- rename(dirca, Cell.Type = cell\_type)  
dirca <- rename(dirca, Observation = observation)  
dirca <- rename(dirca, Primarywall.Thickness.um = primary\_wall\_thickness)  
dirca <- rename(dirca, Secondarywall.Thickness.um = secondary\_wall\_thickness)  
dirca <- rename(dirca, Length.um = length)  
dirca <- rename(dirca, Lumen.Diameter.um = lumen\_diameter)  
dirca <- rename(dirca, Cell.Total.Diameter.um = cell\_total\_diameter)  
dirca <- rename(dirca, Lignified.Y.N = lignified.)  
dirca <- rename(dirca, Number.of.Cell.Width = num\_of\_cell\_width)  
dirca <- rename(dirca, Seriate.Number = seriate\_num)  
  
#Explore dataset  
str(dirca) #check variable class assignment

## 'data.frame': 480 obs. of 13 variables:  
## $ Species : Factor w/ 4 levels "decipiens","mexicana",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Population : Factor w/ 8 levels "eureka","fl",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ Specimen.Number : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ Cell.Type : Factor w/ 5 levels "bark\_fiber","companion\_cell",..: 1 1 1 1 5 5 5 5 3 3 ...  
## $ Observation : int 1 2 3 4 1 2 3 4 1 2 ...  
## $ Primarywall.Thickness.um : int 20 30 40 50 46 56 66 76 26 36 ...  
## $ Secondarywall.Thickness.um: int 30 40 50 60 46 56 66 76 NA NA ...  
## $ Length.um : int 320 330 340 350 1000 1100 1200 1300 160 170 ...  
## $ Lumen.Diameter.um : int 12 22 32 42 240 230 220 210 200 190 ...  
## $ Cell.Total.Diameter.um : int 62 92 122 152 332 342 352 362 226 226 ...  
## $ Lignified.Y.N : Factor w/ 1 level "no": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Number.of.Cell.Width : int NA NA NA NA NA NA NA NA NA NA ...  
## $ Seriate.Number : int NA NA NA NA NA NA NA NA NA NA ...

#Check Variable Class Assignment  
#Correct specimen\_num & observation Class Assignments  
dirca$Specimen.Number <- as.factor(dirca$Specimen.Number)  
dirca$Observation <- as.factor(dirca$Observation)  
str(dirca) #check assignments have changed

## 'data.frame': 480 obs. of 13 variables:  
## $ Species : Factor w/ 4 levels "decipiens","mexicana",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Population : Factor w/ 8 levels "eureka","fl",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ Specimen.Number : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Cell.Type : Factor w/ 5 levels "bark\_fiber","companion\_cell",..: 1 1 1 1 5 5 5 5 3 3 ...  
## $ Observation : Factor w/ 4 levels "1","2","3","4": 1 2 3 4 1 2 3 4 1 2 ...  
## $ Primarywall.Thickness.um : int 20 30 40 50 46 56 66 76 26 36 ...  
## $ Secondarywall.Thickness.um: int 30 40 50 60 46 56 66 76 NA NA ...  
## $ Length.um : int 320 330 340 350 1000 1100 1200 1300 160 170 ...  
## $ Lumen.Diameter.um : int 12 22 32 42 240 230 220 210 200 190 ...  
## $ Cell.Total.Diameter.um : int 62 92 122 152 332 342 352 362 226 226 ...  
## $ Lignified.Y.N : Factor w/ 1 level "no": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Number.of.Cell.Width : int NA NA NA NA NA NA NA NA NA NA ...  
## $ Seriate.Number : int NA NA NA NA NA NA NA NA NA NA ...

#Create tidy database for analysis  
setwd("/Users/Dirca/Desktop/GitHub-RStudio/Dirca Anatomy Mock Data Analysis/\_data/tidy")  
#write.csv(dirca, "\_data/tidy/dirca\_tidy.csv", row.names=F)  
write.csv(dirca, "dirca\_tidy.csv", row.names=F)