data-wrangling

Ata COLAK

2023-07-28

Data Frame Wrangling

Introduction

This report shows a few examples of data frame wrangling using R.

The report shows uses of subsetting and aggregating requested information, from a requested date within a dataset, and showing it using lists and bar plots.

The dataset used for this report is “City of Melbourne’s Trees, with species and dimensions (Urban Forest)”. The dataset can be accessed using the link below:

https://data.melbourne.vic.gov.au/explore/dataset/trees-with-species-and-dimensions-urban-forest/table/

First Part - Loading Dataset

In the first example below, the csv file is read by my computer, and prints the head of the dataset to show that it has been read successfully.

```r
trees <- read.csv("trees.csv")
head(trees, 3)
```

```
##    CoM.ID          Common.Name          Scientific.Name      Genus      Family
## 1 1440992        River red gum Eucalyptus camaldulensis Eucalyptus   Myrtaceae
## 2 1286119        River red gum Eucalyptus camaldulensis Eucalyptus   Myrtaceae
## 3 1439848 European nettle tree         Celtis australis     Celtis Cannabaceae
##   Diameter.Breast.Height Year.Planted Date.Planted Age.Description
## 1                     NA         2009   2009-12-14                
## 2                     80         2008   2008-07-16          Mature
## 3                      4         2009   2009-09-08        Juvenile
##   Useful.Life.Expectency Useful.Life.Expectency.Value Precinct Located.in
## 1                                                  NA       NA       Park
## 2            31-60 years                           60       NA       Park
## 3            31-60 years                           60       NA     Street
##   UploadDate                      CoordinateLocation  Latitude Longitude
## 1 2021-01-10    -37.789042536009, 144.94750113149306 -37.78904  144.9475
## 2 2021-01-10   -37.78989006812276, 144.9256959906416 -37.78989  144.9257
## 3 2021-01-10 -37.795227592098875, 144.91940533967247 -37.79523  144.9194
##    Easting Northing                             geolocation
## 1 319271.4  5815607    -37.789042536009, 144.94750113149306
## 2 317353.2  5815470   -37.78989006812276, 144.9256959906416
## 3 316812.5  5814866 -37.795227592098875, 144.91940533967247
```

Second Part - Subsetting and Aggregating the Trees Dataset to Create a Custom List

In the second part of the code, I have created a combination for the trees of genera asked in the report sheet, to make my job easier, and to make the code easier to read.

I first create a trees subset which only has the genus I will be using.

Secondly, I create a new column which acts as a boolean value, if the tree was planted before 2003, it returns FALSE, if the tree is planted after 2003, it returns TRUE. I called it After2003, and I have used a conditional element selection operator, ifelse, as seen in the code.

Afterwards, I have aggregated the subsetted list by the genera required and the After2003 check. The aggregated data also has a FUN column for the length.

```r
genera <- c("Eucalyptus", "Platanus", "Acacia", "Ulmus", "Pinus")

trees_subset <- subset(trees, Genus %in% genera)

# Create a new column for the planting period
trees_subset$After2003 <- ifelse(trees_subset$Year.Planted <= 2003,
                              "FALSE",
                              "TRUE")


# Aggregate the data by Genus and Period
trees_agg <- aggregate(trees_subset$Year.Planted,
                       by = list(Genus = trees_subset$Genus,
                                 After2003 = trees_subset$After2003),
                       FUN = length)
names(trees_agg)[3] <- "Count"

# Print the result
trees_agg
```

```
##         Genus After2003 Count
## 1      Acacia     FALSE  1386
## 2  Eucalyptus     FALSE  6924
## 3       Pinus     FALSE   173
## 4    Platanus     FALSE  4622
## 5       Ulmus     FALSE  4603
## 6      Acacia      TRUE  5171
## 7  Eucalyptus      TRUE 10384
## 8       Pinus      TRUE   261
## 9    Platanus      TRUE  1065
## 10      Ulmus      TRUE  1324
```

Third Part - Subsetting and Aggregating the Trees Dataset to Create a Custom List and Bar Plot

In the third part of the code, I have created another trees_subset variable trees_subset2, which also has a condition to subset trees planted after 2003.

I have aggregated the subsetted list by genus, and FUN calculates the mean of the diameter breast height, and puts it into the second column.

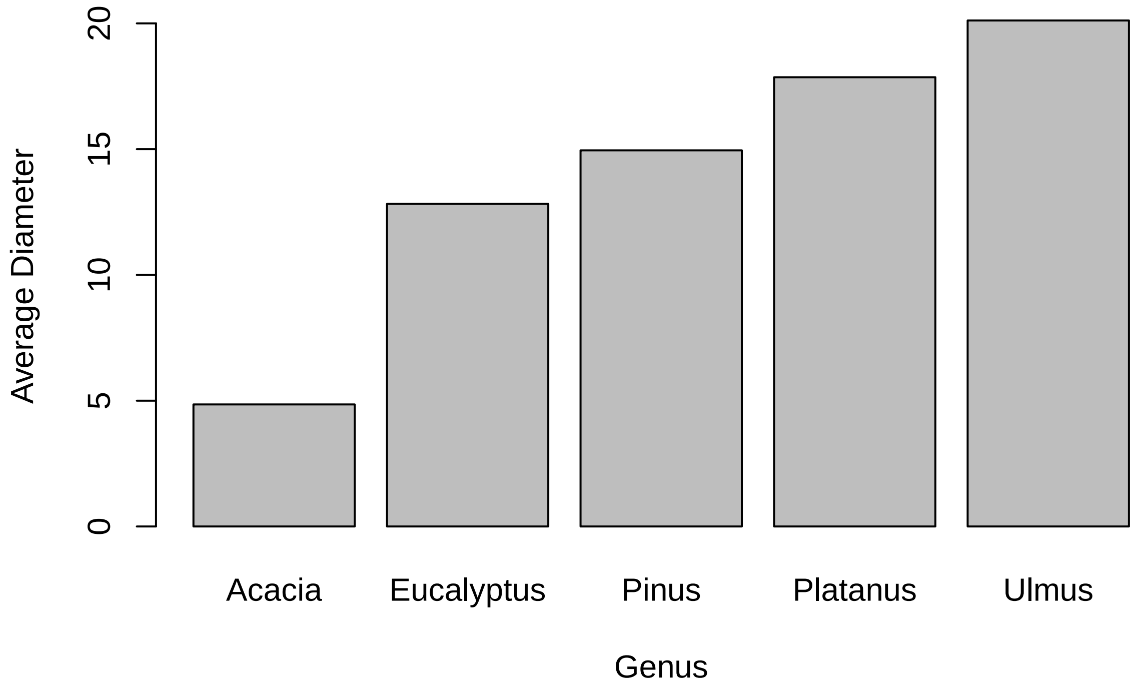
At first, I had problems creating the aggregated list, then I found the solution of giving them names by myself, that fixed the problems within the code.

After the aggregation and naming process, I print the list, and I print the requested bar plot afterwards, with correct labels and data.

```r
trees_subset2 <- subset(trees, Year.Planted >= 2003 & Genus %in% genera)

trees_agg2 <- aggregate(trees_subset2$Diameter.Breast.Height,
                        by=list(trees_subset2$Genus), FUN=mean, na.rm = TRUE)

names(trees_agg2) <- c("Genus", "Diameter")


print(trees_agg2)
```

```
##        Genus  Diameter
## 1     Acacia  4.851351
## 2 Eucalyptus 12.823626
## 3      Pinus 14.952381
## 4   Platanus 17.858473
## 5      Ulmus 20.116022
```

```r
barplot(trees_agg2$Diameter, names.arg = trees_agg2$Genus, xlab = "Genus", ylab = "Average Diameter")
```



Fourth Part - Benefits and Concerns of Sharing Such Data With the Public, and What an AI Engineer Can Do With Them

In my opinion, sharing such detailed data about trees which extends to planting information as old as 1899 is highly beneficial for data analysts, in a way that it allows them to do whatever they want with the data. The examples are limitless, but a few would be:

Analysing data in such way that, by going through the geolocations of the trees, a tree map of Melbourne can be created, which would show how Melbourne's nature was shaped before the trees were cut down,

Analysing pre-dated breast height to compare how the shape changed through the course of over 100 years,

Analysing which trees were more prevelant throughout the years…

The options are limitless, therefore I do not see any negative concerns with sharing this dataset with the public.

Conclusion

In conclusion, this report demonstrated subsetting and aggregating in R for data wrangling, and I believe with my explanations, I proved there are much more different ways of handling this dataset for different purposes.

While doing this report, I have learned how subsetting and aggregation works, and how to analyse datasets to a wider aspect.

What was interesting to me about this report, and the thought provoking question was the amount of possibilities a creative person can see by looking at a big dataset.