Text Analysis HW6

Julia Bright

2025-04-11

```
#read in data
trees <-
↪  read_csv("/Users/julia/Library/CloudStorage/OneDrive-UniversityofNorthCarolinaatChapelHill/Documents
```

Question 1

```
#separate cityname and state in the City column
trees$cityname <- str_match(trees$City, "^[:alpha:]+")[,1]
trees$state <- str_match(trees$City, "[:alpha:]+$")[,1]

#table shows amount of tree records for each state
table(trees$state)
```

```
##
##   AZ   CA   CO   FL   HI   ID   IN   MN   NC   NM   NY   SC   WA
##  827 4062  867  895  918  923  877  760  828  833  831  872  994
```

Question 2

```
#filter dataset to only NC and SC
treesNCSC <- trees %>%
  filter(state== "NC" | state=="SC")
unique(treesNCSC$cityname)
```

```
## [1] "Charleston" "Charlotte"
```

They collected data from Charleston and Charlotte.

Question 3

```
#create genus and species variables by separating the two in the scientific name column
treesNCSC$genus <- str_match(treesNCSC$ScientificName, "([:alpha:]+) ([:alpha:]+)")[,2]
treesNCSC$species <- str_match(treesNCSC$ScientificName, "([:alpha:]+) ([:alpha:]+)")[,3]
```

```r
#find avg canopy size by genus in NC/SC
canopyNCSC <- treesNCSC %>%
  group_by(genus) %>%
  summarize(avg_Cdia = mean(`AvgCdia (m)`))

max(canopyNCSC$avg_Cdia)
```

```
## [1] 13.62316
```

The genus with the largest average crown diameter in North and South Carolina is the Quercus, at 13.6231626 meters.

Question 4

```r
#remove all x's and one space that comes before it to solve problem of hybrid species
↪  denotation
trees$ScientificName <- str_remove_all(trees$ScientificName, "[:space:]x")

#separate genus and species
trees$genus <- str_match(trees$ScientificName, "([:alpha:]+) ([:alpha:]+)")[,2]
trees$species <- str_match(trees$ScientificName, "([:alpha:]+) ([:alpha:]+)")[,3]

#find avg canopy size by genus in the dataset
canopy <- trees %>%
  group_by(genus) %>%
  summarize(avg_cdia = mean(`AvgCdia (m)`))
max(canopy$avg_cdia)
```

```
## [1] 17.70484
```

```r
#group by genus, count the number of distinct species within each genus
totals <- trees %>%
  group_by(genus) %>%
  summarize(num = n_distinct(species))

t1 <- totals[1:42,]
t2 <- totals[43:85,]
kable(list(t1, t2), caption="Number of species within each genus")
```

Table 1: Number of species within each genus

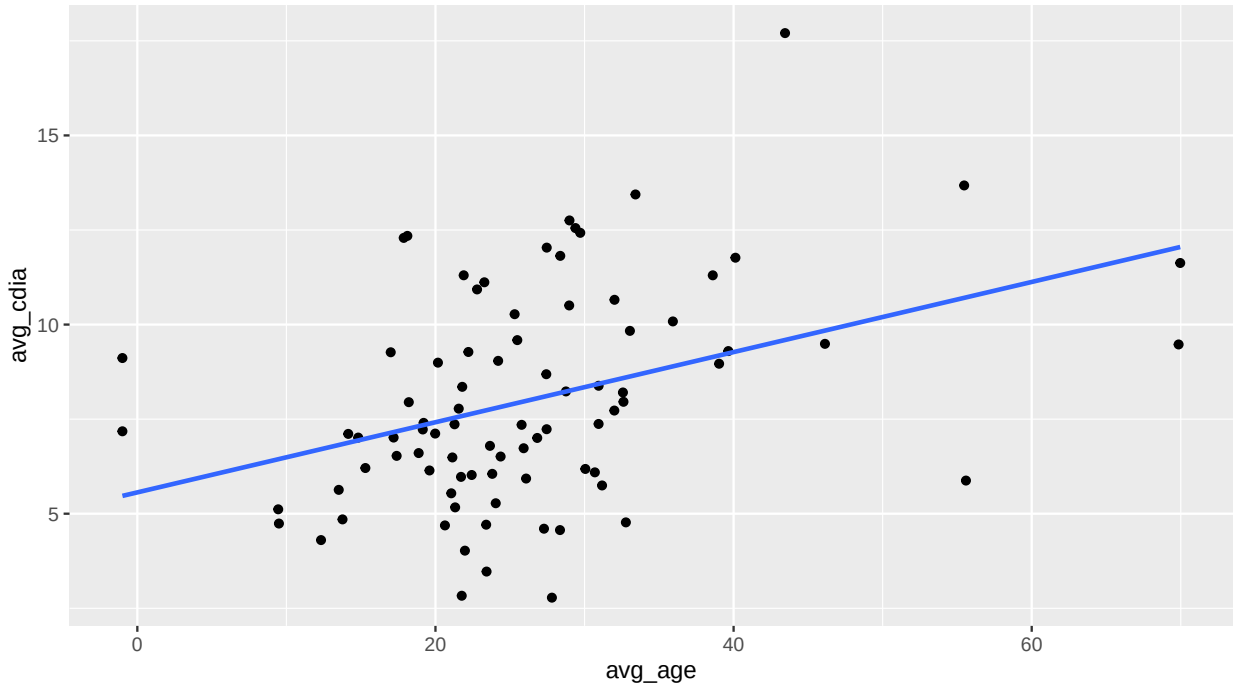
genus	num	genus	num
Acacia	3	Juglans	1
Acer	7	Juniperus	1
Aesculus	1	Koelreuteria	2
Bauhinia	1	Lagerstroemia	3
Betula	2	Liquidambar	1
Brachychiton	1	Liriodendron	1
Butia	1	Magnolia	1
Callistemon	1	Malus	2
Calocedrus	1	Melaleuca	1
Calophyllum	1	Metrosideros	1
Carpinus	1	Morus	2
Carya	1	Olea	1
Cassia	1	Parkinsonia	2
Casuarina	1	Phoenix	2
Catalpa	1	Picea	1
Cedrus	1	Pinus	15
Celtis	3	Pistacia	1
Ceratonia	1	Pittosporum	1
Cercis	1	Platanus	3
Chilopsis	1	Platycladus	1
Cinnamomum	1	Podocarpus	1
Citharexylum	1	Populus	5
Cocos	1	Prosopis	1
Conocarpus	1	Prunus	5
Cordia	1	Pseudotsuga	1
Cornus	1	Pyrus	3
Crataegus	2	Quercus	12
Cupaniopsis	1	Rhus	1
Delonix	1	Robinia	1
Elaeagnus	1	Sabal	1
Elaeodendron	1	Samanea	1
Eriobotrya	1	Schinus	2
Eucalyptus	4	Sequoia	1
Fagus	1	Swietenia	1
Ficus	2	Syagrus	1
Filicium	1	Tabebuia	3
Fraxinus	8	Tilia	3
Ginkgo	1	Triadica	1
Gleditsia	1	Tristaniopsis	1
Gymnocladus	1	Ulmus	4
Ilex	2	Veitchia	1
Jacaranda	1	Washingtonia	2
		Zelkova	1

Extra Credit

```
#group by genus and find average ages
ages <- trees %>%
  group_by(genus) %>%
  summarize(avg_age = mean(Age))

#join canopy averages with age averages
canopyages <- full_join(ages, canopy)

#linear model showing effect of age on canopy size
fit <- lm(avg_cdia ~ avg_age, data=canopyages)
fit
```

```
##
## Call:
## lm(formula = avg_cdia ~ avg_age, data = canopyages)
##
## Coefficients:
## (Intercept)      avg_age
##     5.56374      0.09271
```

```
#scatterplot with trend line of linear model
ggplot(canopyages, aes(x = avg_age, y = avg_cdia)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



There are some significant differences in the average age of the different genera presented in the dataset. I do believe this could be a factor in the results of the average canopy size, given that the genera with the

largest crown diameter in both the NC/SC dataset and the full dataset have an average age of 29 and 43 years, respectively.

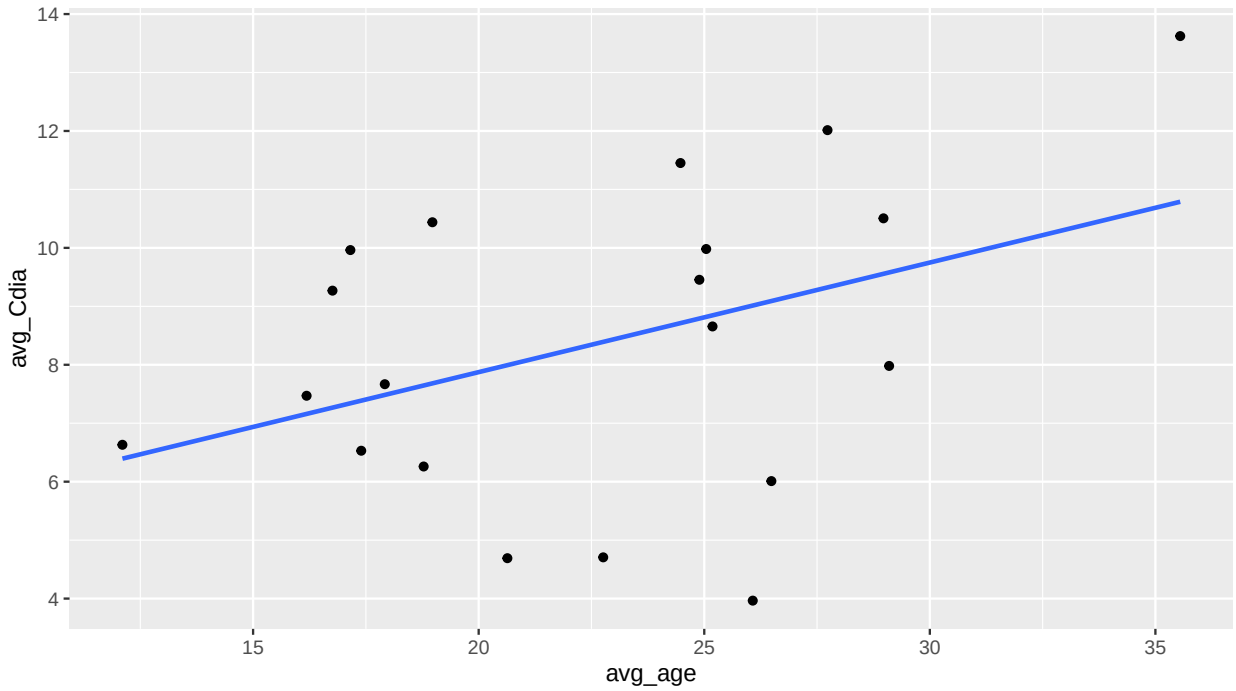
Given this fact, I decided to join together two datasets, both grouped by genus, which summarize the average ages and canopy sizes. Using this joined dataset 'canopyages', I created a linear model to see what effect genus age had on genus crown size, on average. Shown in the scatterplot above is the trend line of that linear regression, revealing that there is an obvious relationship, but that after 40 years the effect is not as strong.

Given that most average crown sizes are below 10 meters, I scanned the scatterplot for a "sweet spot" of tree genera that reach a large crown diameter at a young age. Two stand out, which you may be able to see for your self on the plot, that average around 12.5 meters in diameter and 18 years of age. These two genera are Morus and Zelkova. When sorting the 'canopyages' dataframe by age they stand out among similarly aged genera, as most around their age average a canopy diameter of 6 to 9 meters.

```
#i just saw that I'm supposed to recommend a genera for just NC/SC, I'm leaving in the
↪  above analysis since I already did it

agesNCSC <- treesNCSC %>%
  group_by(genus) %>%
  summarize(avg_age = mean(Age))

canopy_ages_NCSC <- full_join(agesNCSC, canopyNCSC)

#linear model showing effect of age on canopy size
fit2 <- lm(avg_Cdia ~ avg_age, data=canopy_ages_NCSC)
fit2
```

```
##
## Call:
## lm(formula = avg_Cdia ~ avg_age, data = canopy_ages_NCSC)
##
## Coefficients:
## (Intercept)      avg_age
##      4.1250       0.1875
```

```
#scatterplot with trend line of linear model
ggplot(canopy_ages_NCSC, aes(x = avg_age, y = avg_Cdia)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE)
```



Based on this linear model looking only at trees in North and South Carolina, the effect of age on canopy size is even greater than in the full dataset, but this may be due to the smaller sample size. The plot above shows the corresponding trend line. We can see at the top right of the plot our previously identified genus with the largest average crown diameter in the Carolinas, Quercus. There appear to be several candidates for fast-growing trees in the Carolinas. Upon inspection of the dataframe and sorting by age, Betula and Ulmus would be my two recommendations for the goal of planting trees that will provide the greatest canopy cover in the shortest amount of time.