lab2_analysis

AUTHOR
Yazeed Aljohani

PUBLISHED
February 20, 2025

```r
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
```

```r
library(ggplot2)
```

1. Filtering rows

```r
#Question 1: Read in the Minnesota tree growth dataset. Use glimpse to understand the structure and names
# Reading dataset
tree_data <- read.csv("data/tree_dat.csv")
# View structure of dataset
glimpse(tree_data)
```

```
Rows: 131,386
Columns: 8
$ treeID  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ standID <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ stand   <chr> "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A…
```

```
$ year    <int> 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 19…
$ species <chr> "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA"…
$ age     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…
$ inc     <dbl> 0.930, 0.950, 0.985, 0.985, 0.715, 0.840, 0.685, 0.940, 1.165,…
$ rad_ib  <dbl> 10.78145, 11.73145, 12.71645, 13.70145, 14.41645, 15.25645, 15…
```

Answer:

The dataset has 131,386 rows and 8 columns, recording tree growth over time. Each row represents a tree's growth in a specific year.

Main Columns:

- `treeID`: **Unique ID for each tree.**
- `standID`: **Which I believe is a unique identifier for the forest stand.**
- `stand`: **An alphanumeric code representing the stand.**
- `year`: **The year the measurement was taken (from 1897 to 2007).**
- `species`: **Tree species, recorded as a code**
- `age`: **The tree's age at the time of measurement.**
- `inc`: **Growth increment (how much the tree grew in diameter, in mm).**
- `rad_ib`: **Inside bark radius (tree size measurement in mm).**

```r
#Question 2: How many reocrds have been made in stand 1?
stand1_records <- tree_data |> filter(standID == 1)
nrow(stand1_records)
```

```
[1] 979
```

Answer:

979 records were made in stand 1.

```
#Question 3: How many records of the Abies balsamea and Pinus strobus species have been made?
species_count <- tree_data |>
  filter(species %in% c("ABBA", "PIST")) |>
  count(species)
species_count
```

```
  species     n
1    ABBA 13033
2    PIST  4188
```

Answer:

There are 13,033 records for *Abies balsamea* and 4,188 records for *Pinus strobus*.

```
#Question 4: How many trees are older then 200 years old in the last year of the dataset?
max_year <- max(tree_data$year)
old_trees <- tree_data |> filter(year == max_year, age > 200)
nrow(old_trees)
```

```
[1] 7
```

Answer:

7 trees are older than 200 years.

2. Slicing rows

```
#Question 5: What is the oldest tree in the dataset found using slice_max?
oldest_tree <- tree_data |> slice_max(age)
oldest_tree
```

```
  treeID standID stand year species age  inc rad_ib
1     24       2    A2 2007    PIRE 269 0.37 308.84
```

Answer:

The oldest tree in the dataset is a Pinus resinosa (PIRE), recorded in 2007, with an age of 269 years. It had a growth increment of 0.37 mm and an inside bark radius of 308.84 mm.

```r
#Question 6: Find the oldest 5 trees recorded in 2001. Use the help docs to understand optional parameters
oldest_2001 <- tree_data |>
  filter(year == 2001) |>
  slice_max(age, n = 5)
oldest_2001
```

```
  treeID standID stand year species age   inc  rad_ib
1     24       2    A2 2001    PIRE 263 0.210 306.880
2     25       2    A2 2001    PIRE 259 0.280 156.210
3   1595      24    F1 2001    FRNI 212 0.579 156.267
4   1598      24    F1 2001    FRNI 206 0.394 130.251
5   1712      26    F3 2001    FRNI 206 0.168 154.354
```

Answer:

The five oldest trees recorded in 2001 are:

- **Tree ID 24 in stand A2, species Pinus resinosa (PIRE), age 263 years, inside bark radius 306.88 mm.**
- **Tree ID 25 in stand A2, species Pinus resinosa (PIRE), age 259 years, inside bark radius 156.21 mm.**
- **Tree ID 1595 in stand F1, species Fraxinus nigra (FRNI), age 212 years, inside bark radius 156.27 mm.**
- **Tree ID 1598 in stand F1, species Fraxinus nigra (FRNI), age 206 years, inside bark radius 130.25 mm.**
- **Tree ID 1712 in stand F3, species Fraxinus nigra (FRNI), age 206 years, inside bark radius 154.35 mm.**

```r
#Question 7: Using slice_sample, how many trees are in a 30% sample of those recorded in 2002?
sample_2002 <- tree_data |>
  filter(year == 2002) |>
  slice_sample(prop = 0.3)
nrow(sample_2002)
```

```
[1] 687
```

Answer:

A 30% random sample of trees recorded in 2002 includes 687 trees.

3. Arranging Rows

```
# Question 8: Filter all trees in stand 5 in 2007. Sort this subset by descending radius at breast height
stand5_2007 <- tree_data |>
  filter(standID == 5, year == 2007) |>
  arrange(desc(rad_ib)) |>
  slice_head(n = 3)
stand5_2007$treeID
```

```
[1] 128 157 135
```

Answer:

The top three trees in stand 5 in 2007, have the following Tree IDs: 128, 157, and 135

4. Reducing Columns

```
#Question 9: Reduce your full data.frame to [treeID, stand, year, and radius at breast height]. Filter to
smallest_trees <- tree_data |>
  filter(standID == 3, year == 2007) |>
  select(treeID, stand, year, rad_ib) |>
  slice_min(rad_ib, n = 3)

smallest_trees
```

```
  treeID stand year rad_ib
1     50    A4 2007 47.396
```

```
2      56    A4 2007 48.440
3      36    A4 2007 54.925
```

Answer:

The three smallest trees recorded in Stand 3 in 2007, based on inside bark radius:

- **Tree ID 50 in stand A4, with a radius of 47.396 mm.**
- **Tree ID 56 in stand A4, with a radius of 48.440 mm.**
- **Tree ID 36 in stand A4, with a radius of 54.925 mm.**

```r
# Question 10: Use select to remove the stand column. Use glimspe to show the dataset.
modified_data <- tree_data |> select(-stand)
glimpse(modified_data)
```

```
Rows: 131,386
Columns: 7
$ treeID  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ standID <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ year    <int> 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, 19…
$ species <chr> "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA"…
$ age     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18,…
$ inc     <dbl> 0.930, 0.950, 0.985, 0.985, 0.715, 0.840, 0.685, 0.940, 1.165,…
$ rad_ib  <dbl> 10.78145, 11.73145, 12.71645, 13.70145, 14.41645, 15.25645, 15…
```

```r
# Question 11: Look at the help document for dplyr::select and examine the “Overview of selection features
# Question 12: Find a selection pattern that captures all columns with either ‘ID’ or ‘stand’ in the name.
selected_columns <- tree_data |>
  select(matches("ID|stand"))
glimpse(selected_columns)
```

```
Rows: 131,386
Columns: 3
$ treeID  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
```

```
$ standID <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ stand   <chr> "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A…
```

5. Renaming columns

```
#Question 13: Looking back at the data dictionary, rename rad_inc and inc to include _[unit] in the name.

library(dplyr)


tree_data <- read.csv("data/tree_dat.csv")


tree_data <- tree_data |>
  rename(rad_ib_mm = rad_ib, inc_mm = inc)  # Adding _mm to indicate millimeters


glimpse(tree_data)
```

```
Rows: 131,386
Columns: 8
$ treeID    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ standID   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …
$ stand     <chr> "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", …
$ year      <int> 1960, 1961, 1962, 1963, 1964, 1965, 1966, 1967, 1968, 1969, …
$ species   <chr> "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABBA", "ABB…
$ age       <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
$ inc_mm    <dbl> 0.930, 0.950, 0.985, 0.985, 0.715, 0.840, 0.685, 0.940, 1.16…
$ rad_ib_mm <dbl> 10.78145, 11.73145, 12.71645, 13.70145, 14.41645, 15.25645, …
```

6. Creating new columns

```r
# Question 14: A key measurement in forestry in “basal area column”. The metric is computed with the formu
#BA(m2) = 0.00007854⋅DBH^2
#Where DBH is the diameter at breast height (cm). Use mutate to compute DBH in centimeters, and BA in m2 (

library(dplyr)

# Compute DBH (diameter at breast height in cm) and Basal Area (BA in m²)
tree_data <- tree_data |>
  mutate(DBH_cm = (rad_ib_mm * 2) / 10,  # Convert radius (mm) to diameter (cm)
         BA_m2 = 0.00007854 * DBH_cm^2)  # Compute basal area

# Calculate the mean BA for species "POTR" in 2007
mean_BA_POTR_2007 <- tree_data |>
  filter(species == "POTR", year == 2007) |>
  summarize(mean_BA = mean(BA_m2, na.rm = TRUE))

# View the result
mean_BA_POTR_2007
```

```
     mean_BA
1 0.03696619
```

Answer:

The mean basal area (BA) for *Populus tremuloides* (POTR) in 2007 is 0.03697 m^2 (rounded to five decimal places).

```r
#Question 15: Lets say for the sake of our study, trees are not established until they are 5 years of age.
tree_data <- tree_data |>
  mutate(established = if_else(age > 5, TRUE, FALSE))

established_count <- tree_data |> count(established)
established_count
```

```
  established      n
1       FALSE   8883
2        TRUE 122503
```

Answer:

The total number of records from established trees (age > 5) is 122,503, while the number of records from non-established trees (age ≤ 5) is 8,883.

7. case_when / if_else

```
#Question 16: Use mutate and case_when to add a new column to you data.frame that classifies each tree int
tree_data <- tree_data |>
  mutate(DBH_class = case_when(
    DBH_cm < 2.5 ~ "seedling",
    DBH_cm >= 2.5 & DBH_cm < 10 ~ "sapling",
    DBH_cm >= 10 & DBH_cm < 30 ~ "pole",
    TRUE ~ "sawlog"
  ))

class_count_2007 <- tree_data |>
  filter(year == 2007) |>
  count(DBH_class)
class_count_2007
```

```
  DBH_class    n
1      pole 1963
2   sapling  252
3    sawlog   76
```

Answer:

The number of trees in each DBH class recorded in 2007 is:

- **Pole: 1,963 trees**

- **Sapling: 252 trees**
- **Sawlog: 76 trees**

8. Summarizing

```
#Question 17: Compute the mean DBH (in cm) and standard deviation of DBH (in cm) for all trees in 2007. Ex

dbh_stats_2007 <- tree_data |>
  filter(year == 2007) |>
  summarize(mean_DBH = mean(DBH_cm, na.rm = TRUE),
            sd_DBH = sd(DBH_cm, na.rm = TRUE))
dbh_stats_2007
```

```
  mean_DBH   sd_DBH
1 16.09351 6.138643
```

9. Grouped data

```
#Question 18: Compute the per species mean tree age using only those ages recorded in 2003. Identify the t
mean_age_2003 <- tree_data |>
  filter(year == 2003) |>
  group_by(species) |>
  summarize(mean_age = mean(age, na.rm = TRUE)) |>
  arrange(desc(mean_age)) |>
  slice_head(n = 3)
mean_age_2003
```

```
# A tibble: 3 × 2
  species mean_age
  <chr>      <dbl>
1 THOC       127.
```

```
2 FRNI        83.1
3 PIST        73.3
```

Answer:

The three species with the highest mean age in 2003 are:

- **Thuja occidentalis (THOC): 126.64 years**
- **Fraxinus nigra (FRNI): 83.08 years**
- **Pinus strobus (PIST): 73.29 years**

10. Counting

```
#Question 19: In a single summarize call, find the number of unique years with records in the data set alc
year_stats <- tree_data |>
  summarize(unique_years = n_distinct(year),
            first_year = min(year),
            last_year = max(year))
year_stats
```

```
  unique_years first_year last_year
1          111       1897      2007
```

Answer:

The dataset contains **111 unique years** of recorded tree growth data.
The **first recorded year** in the dataset is **1897**.

```
#Question 20: Determine the stands with the largest number of unique years recorded. Report all stands wit
stand_years <- tree_data |>
  group_by(standID) |>
  summarize(unique_years = n_distinct(year)) |>
```

```
  filter(unique_years == max(unique_years))
stand_years
```

```
# A tibble: 5 × 2
  standID unique_years
    <int>        <int>
1       1          111
2      15          111
3      16          111
4      17          111
5      24          111
```

Answer:

The **stands with the largest number of unique years recorded (111 years)** are:

- **Stand ID 1**
- **Stand ID 15**
- **Stand ID 16**
- **Stand ID 17**
- **Stand ID 24**

Final Question:

```
#Use a combination of dplyr verbs to compute these values and report the 3 species with the fastest growth

#Lastly, find and include an image of the fastest growing species. Add the image to your images directory.

growth_rates <- tree_data |>
  group_by(treeID, species) |>
  arrange(year) |>
  mutate(annual_growth = diff(c(NA, DBH_cm))) |>
```

```
  group_by(species) |>
  summarize(mean_growth = mean(annual_growth, na.rm = TRUE),
            sd_growth = sd(annual_growth, na.rm = TRUE)) |>
  arrange(desc(mean_growth))

fastest_species <- growth_rates |> slice_head(n = 3)
slowest_species <- growth_rates |> slice_tail(n = 3)
list(fastest_species, slowest_species)
```

```
[[1]]
# A tibble: 3 × 3
  species mean_growth sd_growth
  <chr>         <dbl>     <dbl>
1 PIRE          0.358     0.258
2 POTR          0.331     0.218
3 PIBA          0.326     0.247
```

```
[[2]]
# A tibble: 3 × 3
  species mean_growth sd_growth
  <chr>         <dbl>     <dbl>
1 QURU          0.168    0.0869
2 THOC          0.153    0.0909
3 LALA          0.150    0.113
```

Answer:

Fastest Growing Species: Quercus rubra

The fastest-growing species in this dataset is **Quercus rubra (Red Oak)**, with an average annual growth rate of **0.1675 mm**.

Image of Quercus rubra

Quercus rubra - Fastest Growing Tree