Lab Program - 5

R V Abhishek

2025-09-16

Advanced Data Manipulation with dplyr and Complex Grouping

Objective - The goal of this program is to test advanced data manipulation techniques using the dplyr package.

```
# Load necessary libraries
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
```

```
library(nycflights13)
library(ggplot2)
library(zoo)
```

```
Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```
# Preview the Star Wars Dataset
data("starwars")
head(starwars)
```

```
# A tibble: 6 x 14
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>           <dbl> <chr> <chr>
1 Luke Sky~    172    77 blond      fair       blue             19   male  mascu~
```

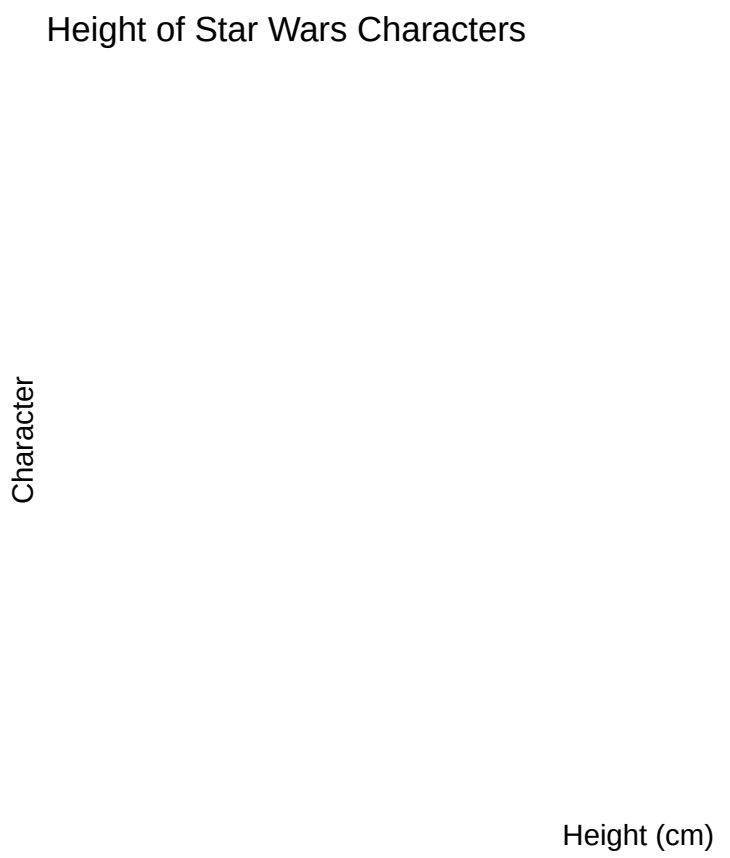
```
2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu~
3 R2-D2         96    32 <NA>       white, bl~ red             33   none  mascu~
4 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
5 Leia Org~    150    49 brown      light      brown           19   fema~ femin~
6 Owen Lars    178   120 brown, gr~ light      blue            52   male  mascu~
# i 5 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>
```

```r
# Selct specific columns(name, species, height, mass), and filtering out the missing species and arrang
starwars_filtered <- starwars %>%
  select(name, species, height, mass) %>%
  filter(!is.na(species) & is.na(height) & height > 100) %>%
  arrange(desc(height))

#Display the filtered data
head(starwars_filtered)
```

```
# A tibble: 0 x 4
# i 4 variables: name <chr>, species <chr>, height <int>, mass <dbl>
```

```r
# Plotting the filtered data
ggplot(starwars_filtered, aes(x = reorder(name, -height), y = height, fill = species)) +
  geom_point(stat = "identity") +
  coord_flip() +
  labs(title = "Height of Star Wars Characters",
       x = "Character",
       y = "Height (cm)") +
  theme_minimal()
```



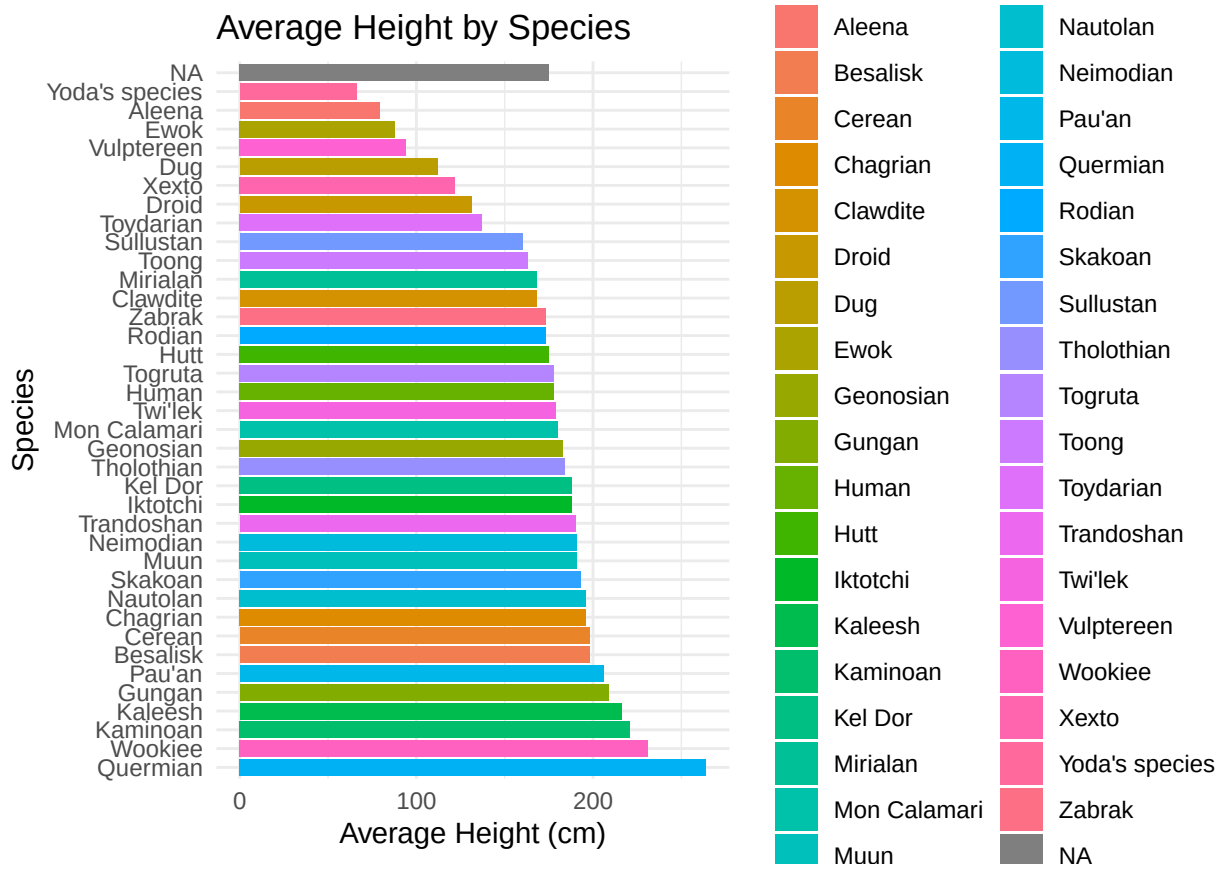

```
# Grouping by species, calculating average height and mass, and counting observation
species_summary <- starwars %>%
  group_by(species) %>%
  summarise(
    avg_height = mean(height, na.rm = TRUE),
    avg_mass = mean(mass, na.rm = TRUE),
    count = n()
  ) %>%
  arrange(desc(count))

# Display the species summary
head(species_summary)
```

```
# A tibble: 6 x 4
  species  avg_height avg_mass count
  <chr>         <dbl>    <dbl> <int>
1 Human          178      81.3    35
2 Droid          131.     69.8     6
3 <NA>           175      81       4
4 Gungan         209.     74       3
5 Kaminoan       221      88       2
6 Mirialan       168      53.1     2
```

```
# Plotting the average height
ggplot(species_summary, aes(x = reorder(species, -avg_height), y = avg_height, fill = species)) +
```

```
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Average Height by Species",
       x = "Species",
       y = "Average Height (cm)") +
  theme_minimal()
```

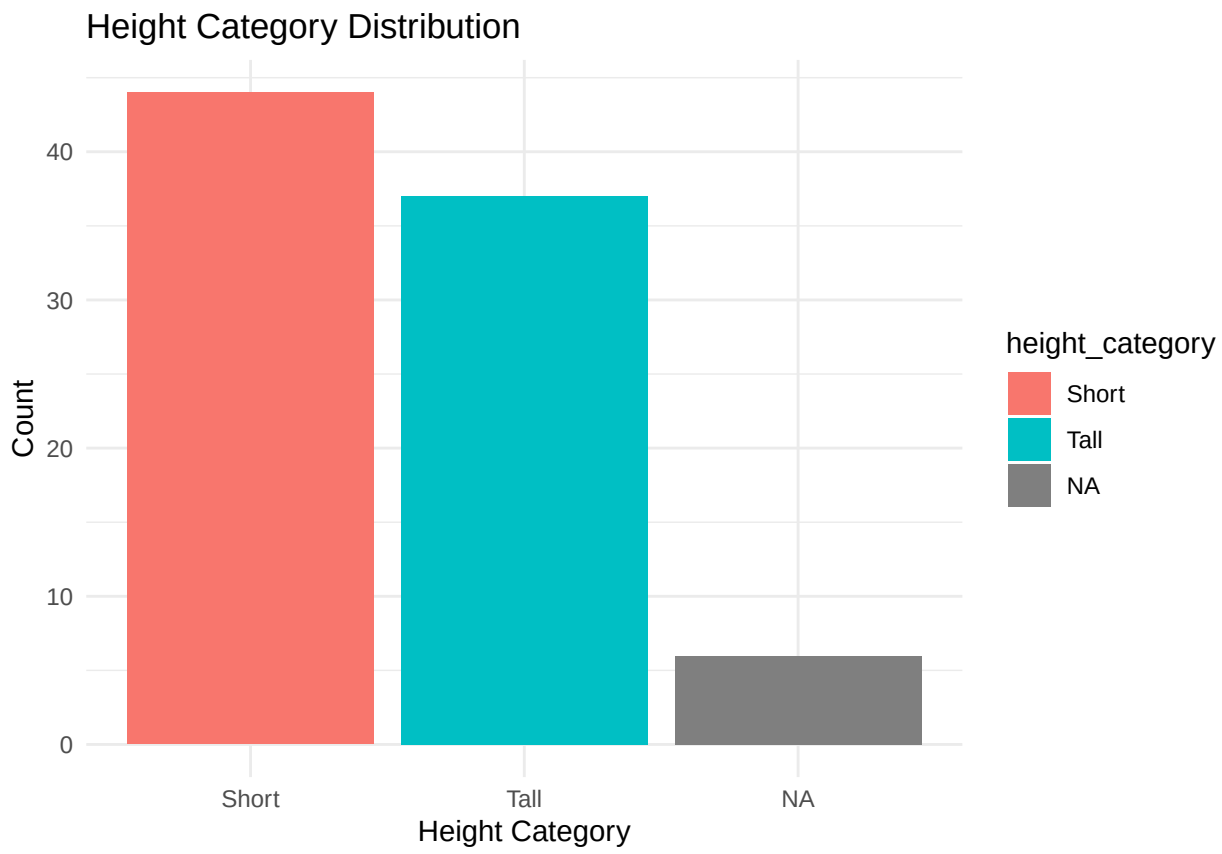



```
# Adding a new column that classifies characters based on height
starwars_classified <- starwars %>%
  mutate(height_category = ifelse(height < 180, "Tall", "Short"))

# Display the classified data
head(starwars_classified)
```

```
# A tibble: 6 x 15
  name      height  mass hair_color skin_color eye_color birth_year sex   gender
  <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
2 C-3PO        167    75 <NA>       gold       yellow         112   none  mascu~
3 R2-D2         96    32 <NA>       white, bl~ red             33   none  mascu~
4 Darth Va~    202   136 none       white      yellow          41.9 male  mascu~
5 Leia Org~    150    49 brown      light      brown           19   fema~ femin~
6 Owen Lars    178   120 brown, gr~ light      blue            52   male  mascu~
# i 6 more variables: homeworld <chr>, species <chr>, films <list>,
#   vehicles <list>, starships <list>, height_category <chr>
```

```r
# Plotting height Category distribution
ggplot(starwars_classified, aes(x = height_category, fill = height_category)) +
  geom_bar() +
  labs(title = "Height Category Distribution",
       x = "Height Category",
       y = "Count") +
  theme_minimal()
```




```r
# Joining with another dataset (flights dataset from nycflights13)
data("flights")
data("airlines")

# Inner join flights with airlines on the common column "carrier"
flights_inner_join <- flights %>%
  inner_join(airlines, by = "carrier")

# Outer join flights with airlines on the common column "carrier"
flights_outer_join <- flights %>%
  full_join(airlines, by = "carrier")

# Display the joined data
head(flights_inner_join)
```

```
# A tibble: 6 x 20
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
```

```
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
1  2013     1     1      517            515         2      830            819
2  2013     1     1      533            529         4      850            830
3  2013     1     1      542            540         2      923            850
4  2013     1     1      544            545        -1     1004           1022
5  2013     1     1      554            600        -6      812            837
6  2013     1     1      554            558        -4      740            728
# i 12 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>, name <chr>
```

```r
head(flights_outer_join)
```

```
# A tibble: 6 x 20
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
1  2013     1     1      517            515         2      830            819
2  2013     1     1      533            529         4      850            830
3  2013     1     1      542            540         2      923            850
4  2013     1     1      544            545        -1     1004           1022
5  2013     1     1      554            600        -6      812            837
6  2013     1     1      554            558        -4      740            728
# i 12 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>, name <chr>
```

```r
# Calculating a 5 period rolling average of arrival delays and cumulative sum
flights_rolling <- flights %>%
  arrange(year, month, day) %>%
  mutate(
    rolling_avg_delay = zoo::rollmean(arr_delay, 5, fill = NA),
    cumulative_delay = cumsum(arr_delay)
  )

# Compute the rolling average and cumulative suma
flights_rolling <- flights %>%
  arrange(year, month, day) %>%
  mutate(
    arr_delay = ifelse(is.na(arr_delay), 0, arr_delay),
    rolling_avg_delay = rollmean(arr_delay, 5, fill = NA),
    cumulative_delay = cumsum(arr_delay)
  )

# Display the transformed data
head(flights_rolling)
```

```
# A tibble: 6 x 21
   year month   day dep_time sched_dep_time dep_delay arr_time sched_arr_time
  <int> <int> <int>    <int>          <int>     <dbl>    <int>          <int>
1  2013     1     1      517            515         2      830            819
2  2013     1     1      533            529         4      850            830
3  2013     1     1      542            540         2      923            850
4  2013     1     1      544            545        -1     1004           1022
```

```
5  2013     1     1      554            600        -6      812            837
6  2013     1     1      554            558        -4      740            728
# i 13 more variables: arr_delay <dbl>, carrier <chr>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dttm>, rolling_avg_delay <dbl>,
#   cumulative_delay <dbl>
```

```r
# Plotting the rolling average and cumulative delays
ggplot(flights_rolling, aes(x = day)) +
  geom_line(aes(y = rolling_avg_delay, color = "Rolling Average Delay")) +
  geom_line(aes(y = cumulative_delay / 1000, color = "Cumulative Delay (x1000)")) +
  labs(title = "Rolling Average and Cumulative Arrival Delays",
       x = "Day of the Month",
       y = "Delay (minutes)") +
  scale_color_manual(values = c("Rolling Average Delay" = "blue", "Cumulative Delay (x1000)" = "red")) +
  theme_minimal()
```

```
Warning: Removed 4 rows containing missing values or values outside the scale range
(`geom_line()`).
```