Lab 5

Enter the names of the group members here: Austine Do, Graceanne Becker, Catherine Zhong

This assignment is due by the end of the lab. Only one student in the group submits a pdf file on Gradescope.

For all questions, include the R commands/functions that you used to find your answer (show R chunk). Answers without supporting code will not receive credit. Write full sentences to describe your findings.

In this lab, you will explore the dataset `starwars` which comes with `dplyr`. Let's first load the packages we will need to complete this lab (`dplyr` and `ggplot2`, all contained in`tidyverse`):

```
# Load the package
library(tidyverse)
```

Take a quick look at the dataset:

```
# Take a quick look
head(starwars)
```

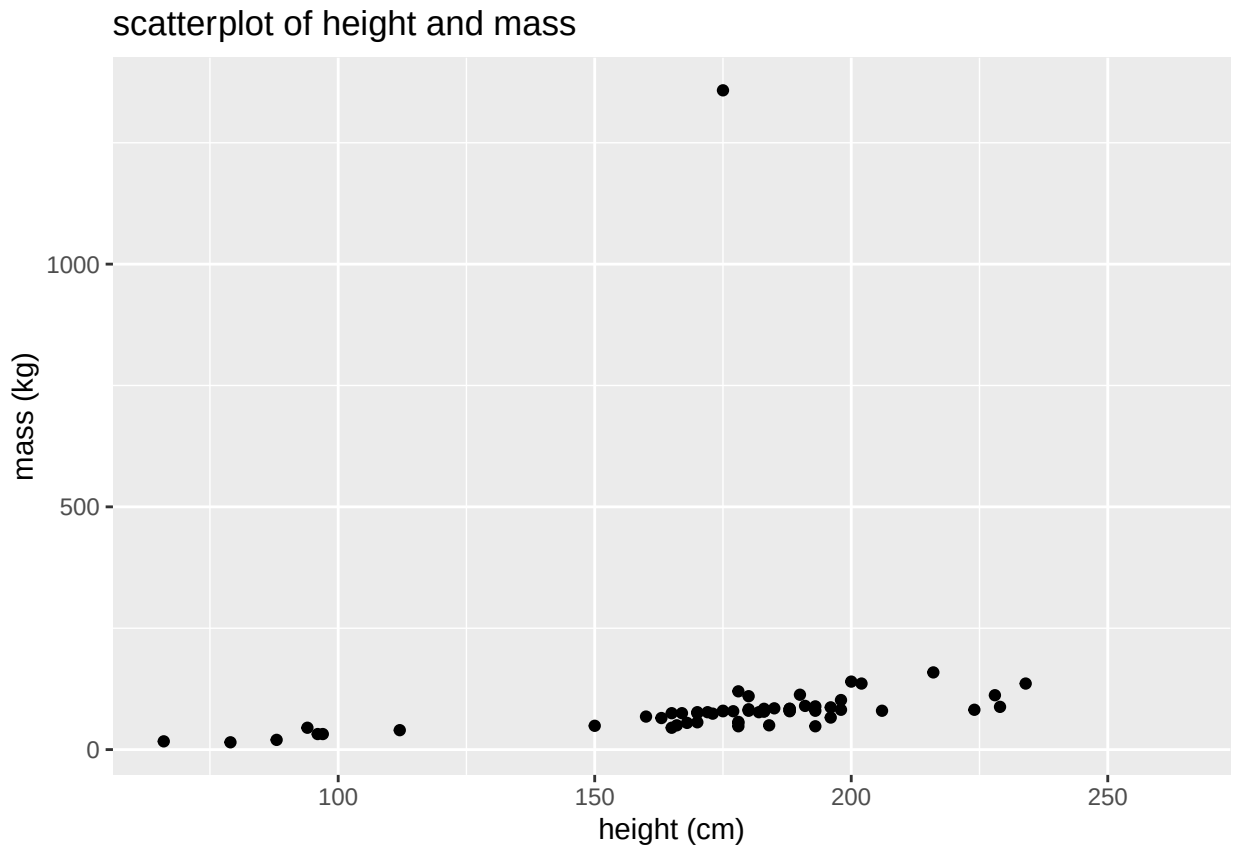
```
## # A tibble: 6 x 14
##   name      height  mass hair_color skin_color eye_color birth_year sex   gender
##   <chr>      <int> <dbl> <chr>      <chr>      <chr>          <dbl> <chr> <chr>
## 1 Luke Sky~    172    77 blond      fair       blue            19   male  mascu~
## 2 C-3PO        167    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>
```

This dataset contains information about Starwars characters which we will investigate using `dplyr` functions.

Question 1: (4 pts)

Using `ggplot`, represent the relationship between `height` and `mass` (make sure to include units in the labels: look at the documentation). Do you notice anything is that visualization?

```
# Scatterplot of heigth and mass of all the Star Wars character in the dataset
starwars |>
    ggplot() +
        geom_point(aes(x = height, y = mass)) +
        labs(title = 'scatterplot of height and mass',
             x = 'height (cm)',
             y = 'mass (kg)')
```



We noticed that there was one outlier in terms of mass at around 1250-1500 kg

---

Question 2: (4 pts)

Using `dplyr` core functions, create a new variable to calculate the Body Mass Index (BMI) for a height in meters and a weight in kilograms:

$$BMI = \frac{weight}{height^2}$$

Only display the top 5 observations for BMI and only keep relevant information (`name`, `species`, `height`, `mass`). Who has the highest BMI in the dataset?

```
# Calculate BMI of all character and display the characters with the highest BMI
starwars |>
    mutate(BMI = mass / ((height/100) ** 2)) |>
    slice_max(n = 5, BMI)
```

```
## # A tibble: 5 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 Jabba De~    175  1358 <NA>       green-tan~ orange           600 herm~ mascu~
```

```
## 2 Dud Bolt      94    45 none       blue, grey yellow           NA male  mascu~
## 3 Yoda          66    17 white      green      brown           896 male  mascu~
## 4 Owen Lars    178   120 brown, gr~ light      blue             52 male  mascu~
## 5 IG-88        200   140 none       metal      red              15 none  mascu~
## # i 6 more variables: homeworld <chr>, species <chr>, films <list>,
## #   vehicles <list>, starships <list>, BMI <dbl>
```

Jabba Desilijic Tiure has the highest BMI in the dataset.

Question 3: (4 pts)

Using `dplyr` core functions, find how many characters there are *per species*. What are the two most common species?

```
# Display the counts of each species and arrange them from highest to lowest count
starwars |>
    group_by(species) |>
    summarize(character_per_species = sum(n())) |>
    arrange(desc(character_per_species))
```

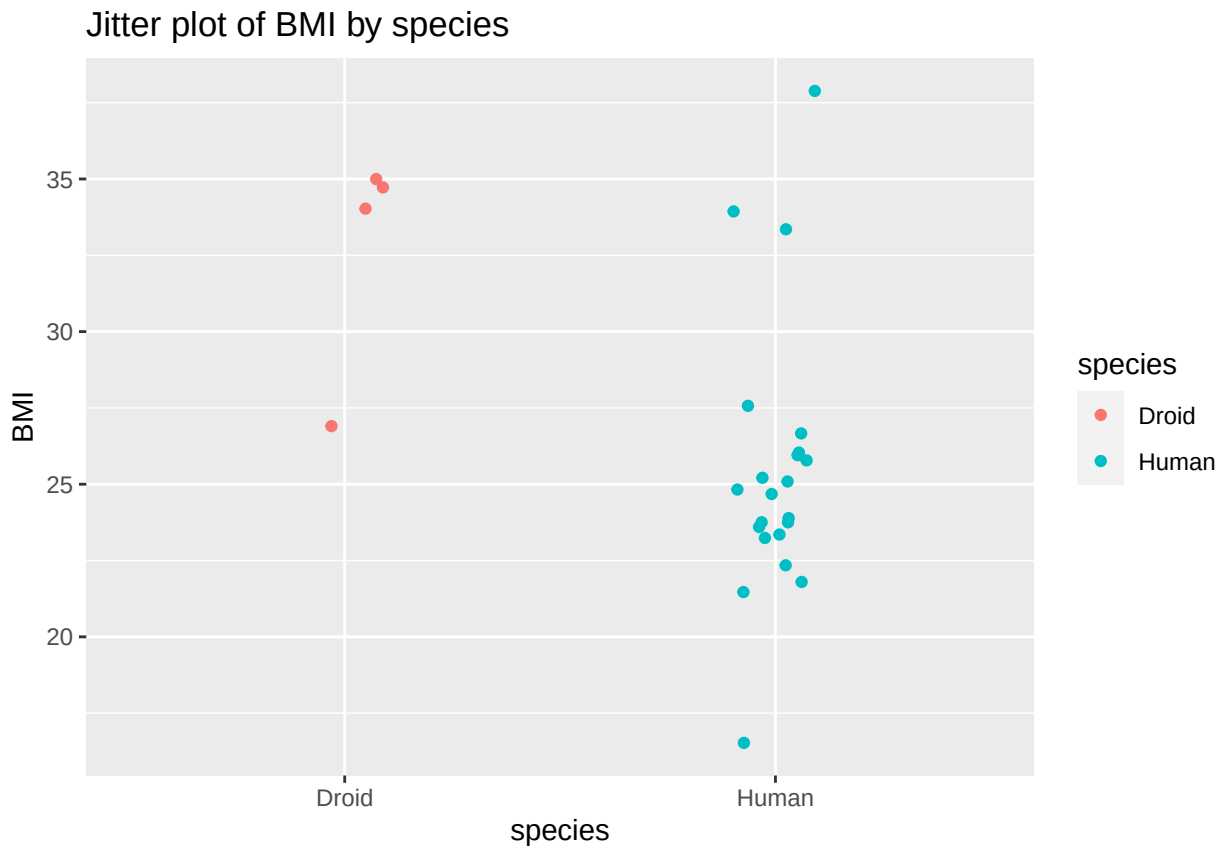
```
## # A tibble: 38 x 2
##    species  character_per_species
##    <chr>                    <int>
##  1 Human                       35
##  2 Droid                        6
##  3 <NA>                         4
##  4 Gungan                       3
##  5 Kaminoan                     2
##  6 Mirialan                     2
##  7 Twi'lek                      2
##  8 Wookiee                      2
##  9 Zabrak                       2
## 10 Aleena                       1
## # i 28 more rows
```

The two most common species were human at 35 characters and droid at 6 characters.

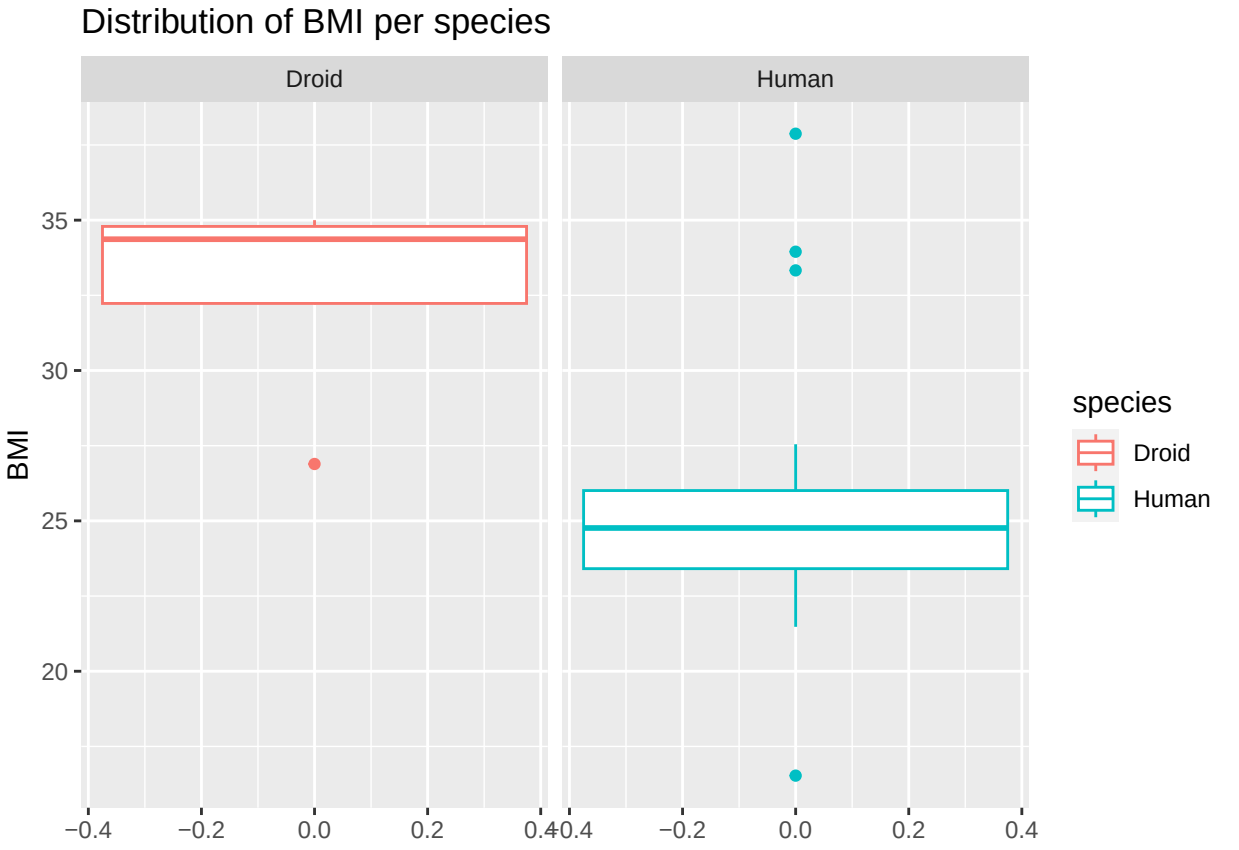
Question 4: (5 pts)

Using `dplyr` core functions and `ggplot`, compare the distribution of BMI between the two species found in the previous question. Use `geom_boxplot()` to compare the two species but also show the data with `geom_jitter(width = 0.1)`. Which of the two species seem to have the highest BMI on average? Is that reasonable to make such a comparison?

```r
#
starwars |>
    mutate(BMI = mass / ((height/100) ** 2)) |>
    filter(species %in% c('Human', 'Droid')) |>
    ggplot() +
        geom_jitter(aes(x = species, y = BMI, color = species), width = 0.1) +
        labs(title = 'Jitter plot of BMI by species',
             x = 'species',
             y = 'BMI')
```




```r
starwars |>
    mutate(BMI = mass / ((height/100) ** 2)) |>
    filter(species %in% c('Human', 'Droid')) |>
    ggplot() +
        geom_boxplot(aes(y = BMI, color = species)) +
        facet_wrap(~species) +
        labs(title = 'Distribution of BMI per species',
             y = 'BMI')
```



It seems that Drone has the highest average BMI out of the 2 species we compared but it does not seem like that is a fair comparison to make since there are many more human observations than drone observations in the dataset

---

Question 5: (6 pts)

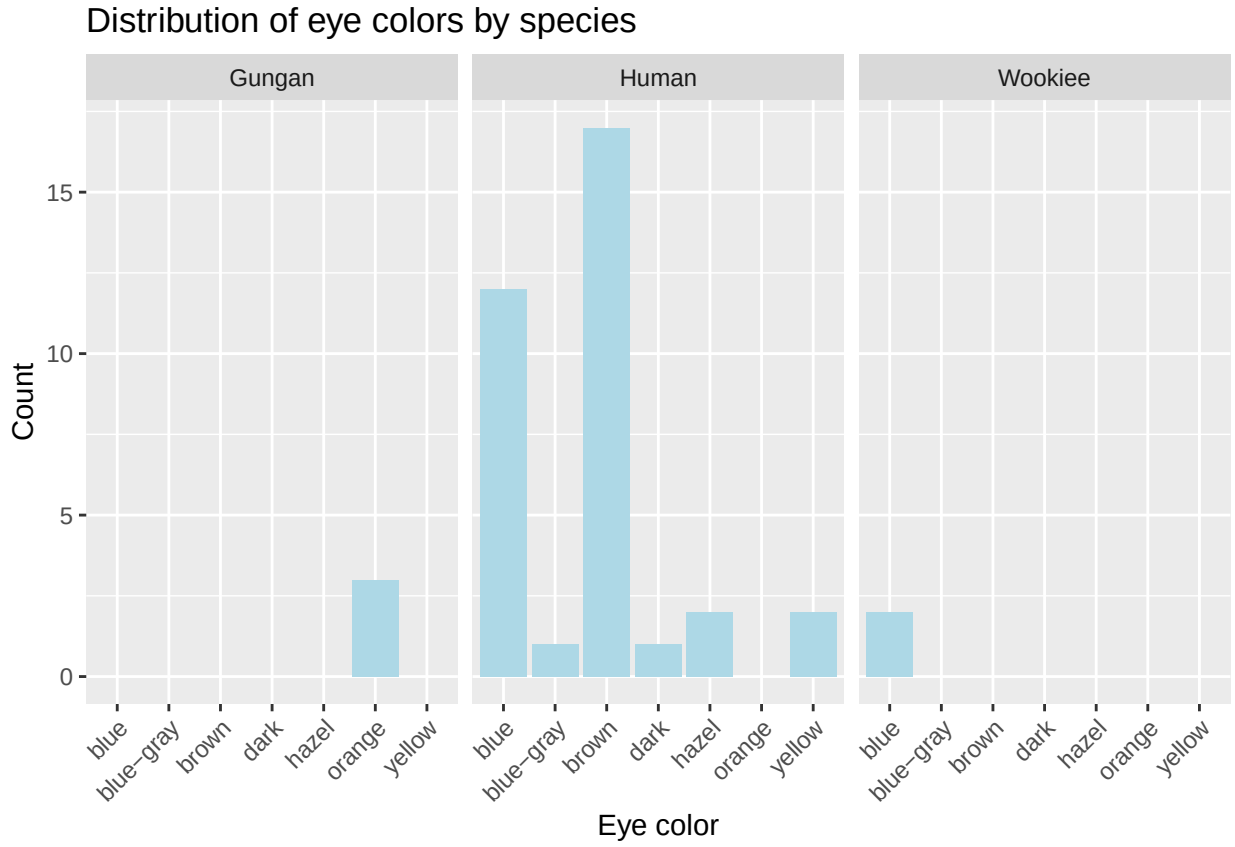
Investigate some other features of the Starwars characters. Write a research question to explore two variables about the Starwars characters (excluding `films`, `vehicles`, and `starships`, we haven't learned how to deal with these types of variables yet!). *For example, (create a question of your own, don't use this one!): How does hair color vary across homeworlds?*

How does eye color vary across Humans, Gungans, and Wookiees?

Answer your research question using some `dplyr` functions (to find summary statistics for example) and a `ggplot` visualization. Include a title to your viz and interpret what you see!

```
#
starwars |>
    filter(species %in% c('Human', 'Gungan', 'Wookiee')) |>
    group_by(species) |>
    ggplot() +
        geom_bar(aes(x = eye_color), fill = 'lightblue') +
        theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
```

```
        labs(title = 'Distribution of eye colors by species',
             x = 'Eye color',
             y = 'Count') +
        facet_wrap(~species)
```




```
filtered_starwars <- starwars |>
                      filter(species %in% c('Human','Gungan','Wookiee'))
table(filtered_starwars$species, filtered_starwars$eye_color)
```

```
##
##           blue blue-gray brown dark hazel orange yellow
##   Gungan     0         0     0    0     0      3      0
##   Human     12         1    17    1     2      0      2
##   Wookiee    2         0     0    0     0      0      0
```

We see from the visualization that both the Gungans and Wookiees have only one color for the species while the Humans have the largest range of eye colors.

Question 6: (1 pt)

After investigating some characterisitcs of Starwars characters, did the data match your expectations or not? If the data differed from your expectation, provide a possible explanation for why the data differed from what you expected.

The data didn't match our expectations because we thought there would be more variety in eye color in all the species we explored in the visualization but we only saw that humans had a wide variety of eye colors.

Formatting: (1 pt)

Make sure the names of all group members are included at the beginning of the document.

Knit your file! You can knit into pdf directly or into html. Once it knits in html, click on `Open in Browser` at the top left of the window pops out. Print your html file into pdf from your browser.

Any issue? Ask other classmates or TA!

Finally, remember to select pages for each question when submitting your pdf to Gradescope and to identify your group members.