Assignment 2

Junn Pham, 300538618

2023-05-04

When submitting my work I confirm:

- I have completed all steps of the attached assessment on my own,
- I have not used any unauthorised materials while completing this assessment, and
- I have not given anyone else access to my assessment.

Question 1

a.

```
movies<- read.csv("movies500.csv", header = TRUE)
movies_genres<- read.csv("movies500_genres.csv", header = TRUE)
genres<- read.csv("genres.csv", header = TRUE)
```

b.

```
library(DBI)
library(RSQLite)
if (file.exists("movies.sqlite")) {
  unlink('movies.sqlite', recursive=TRUE)
}
movies_conn <- dbConnect(RSQLite::SQLite(), "movies.sqlite")
```

c.

```
dbWriteTable(movies_conn, "movies", movies, overwrite=TRUE)
dbWriteTable(movies_conn, "movies_genres", movies_genres, overwirte=TRUE)
```

d.

```
SELECT count(*) as 'number of rows' FROM movies
```

Table 1: 1 records

number of rows
500

e.

```
SELECT title, runtime, release_date FROM movies
WHERE runtime > 480
ORDER BY runtime asc
```

Table 2: 8 records

title	runtime	release_date
Planet Earth	550	2006-12-10
Tie Xi Qu: West of the Tracks	551	2002-04-26
Shoah	566	1985-11-01
The Godfather Trilogy: 1972-1990	583	1992-10-17
New York: A Documentary Film	600	1999-11-14
The Civil War	680	1990-09-23
The Story of Film: An Odyssey	900	2011-09-03
Heimat: A Chronicle of Germany	925	1984-09-16

f.

```
SELECT title FROM movies
WHERE title LIKE '%love%'
```

Table 3: Displaying records 1 - 10

title
Marvin Hamlisch: What He Did For Love
Love at 16
My Future Love
Frankie Boyle: Hurt Like You've Never Been Loved
Harold and Lillian: A Hollywood Love Story
Leather Jacket Love Story
Love Torn in a Dream
The Loves of Pharaoh
Love You You
From Mexico With Love

g.

```
CREATE TABLE genres (
  genre_id integer,
  genre_name text
)
```

h.

```
dbWriteTable(movies_conn, "genres", genres, append=TRUE)
```

i.

```
INSERT INTO genres(genre_id, genre_name)
VALUES (3579,"University Comedy")
```

j.

```
UPDATE genres
SET genre_name="University Tragedy"
WHERE genre_id=3579
```

k.

```
SELECT movies_genres.genre_id from movies
JOIN movies_genres ON movies.tmdbId = movies_genres.tmdbId
WHERE movies.title = "Running Wild"
```

Table 4: 2 records

genre_id
12
18

l.

```
SELECT genres.genre_name from genres
JOIN (
  SELECT * from movies
  JOIN movies_genres ON movies.tmdbId = movies_genres.tmdbId
  WHERE movies.title = "Running Wild"
) X
ON X.genre_id = genres.genre_id
```

Table 5: 2 records

genre_name
Adventure
Drama

m.

```
SELECT genres.genre_name,count(*) from genres
LEFT JOIN (
  SELECT * from movies
  JOIN movies_genres ON movies.tmdbId = movies_genres.tmdbId
) X
ON X.genre_id = genres.genre_id
GROUP BY genres.genre_id
HAVING count(*) >= 20
ORDER BY 2 DESC
```

Table 6: 7 records

genre_name	count(*)
Drama	164
Documentary	146
Comedy	113
Romance	40
Music	28
Action	22
Crime	20

```
dbDisconnect(movies_conn)
```

Question 2

```
library(dplyr)
```

```
library(tidyr)
vehicles <- read.csv("motor_vehicle_modified.csv", stringsAsFactors = FALSE)
vehicles_R <- vehicles
```

a.

```
##dplyr
summarise(filter(vehicles, transmission_type=='4-gear auto' & (make=='Kia' | make=='Honda')),count=n())
```

```
##   count
## 1    13
```

```
##Base R
sum(vehicles_R$transmission_type=='4-gear auto' & (vehicles_R$make=='Kia' | vehicles_R$make=='Honda'))
```

```
## [1] 13
```

b.

```
##dplyr
vehicles <- vehicles %>% select(-vehicle_usage, -vehicle_type)

##Base R
vehicles_R <- subset(vehicles_R, select = -c(vehicle_usage, vehicle_type))
```

c.

```
##dplyr
vehicles_country_status <- vehicles %>% group_by(original_country, import_status) %>%
                           summarise(Count=n())
```

```
## `summarise()` has grouped output by 'original_country'. You can override using
## the `.groups` argument.
```

```
##Base R
vehicles_country_status_R <- vehicles_R
vehicles_country_status_R <- aggregate(vehicles_country_status_R$original_country, by=list(vehicles_cou
colnames(vehicles_country_status_R) <- c("original_country", "import_status", "Count")
```

d.

```
##dplyr
sorted_vehicles_country_status <- filter(vehicles_country_status, import_status=='used') %>% arrange(de
sorted_vehicles_country_status[1:3,]
```

```
## # A tibble: 3 x 3
## # Groups:   original_country [3]
##   original_country import_status Count
##   <chr>            <chr>         <int>
## 1 Japan            used           1172
## 2 Germany          used            137
## 3 United Kingdom   used             34
```

```
##Base R
sorted_vehicles_country_status_R <- vehicles_country_status_R[vehicles_country_status_R$import_status==
sorted_vehicles_country_status_R <- sorted_vehicles_country_status_R[order(sorted_vehicles_country_statu
head(sorted_vehicles_country_status_R,3)
```

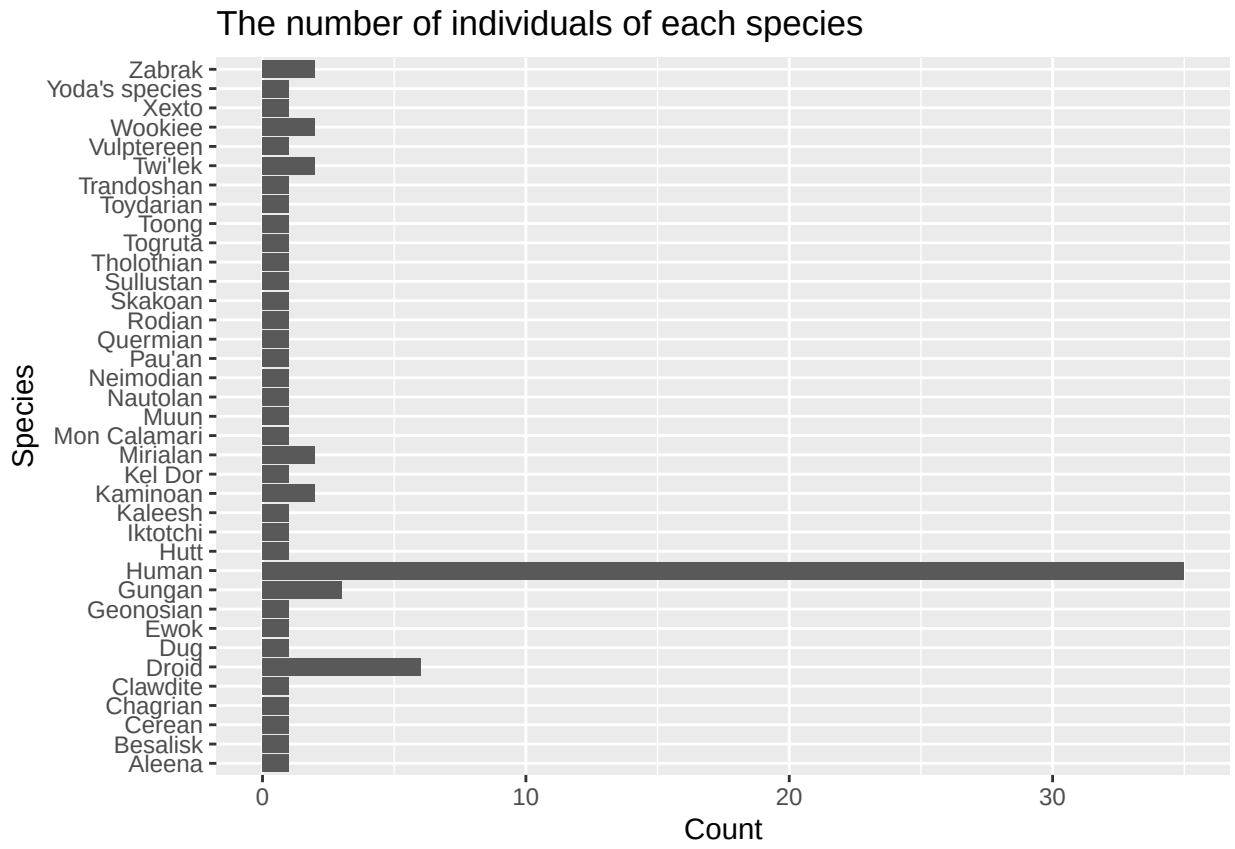
```
##    original_country import_status Count
## 42            Japan          used  1172
## 40          Germany          used   137
## 48   United Kingdom          used    34
```

Question 3

```
library(ggplot2)
data(starwars)
```

a.

```
filter(starwars, is.na(species)==FALSE) %>% ggplot() +
  geom_bar(aes(x=species)) +
  coord_flip() +
  labs(x = "Species", y = "Count", title = "The number of individuals of each species")
```



b.

```
starwars_species <- starwars %>% group_by(species) %>% summarise(num=n())
starwars <- full_join(starwars,starwars_species)
```

```
## Joining with `by = join_by(species)`
```

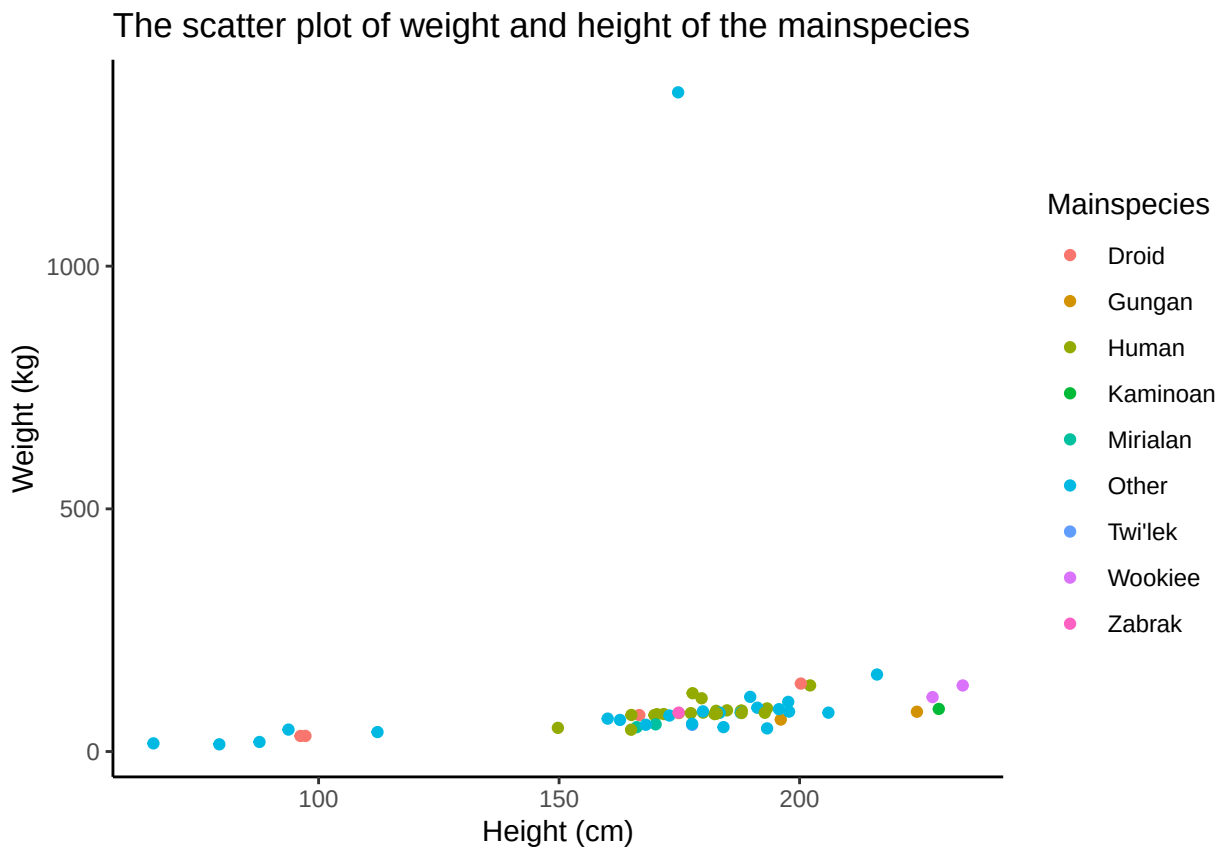
c.

```
starwars <- starwars %>% mutate(Mainspecies=starwars$species) %>%
            rows_update(tibble(num=1, Mainspecies="Other"))
```

```
## Matching, by = "num"
```

d.

```
plot.settings <- filter(starwars, is.na(Mainspecies)==FALSE & is.na(height)==FALSE & is.na(mass)==FALSE
  labs(x = "Height (cm)", y = "Weight (kg)", title = "The scatter plot of weight and height of the mains
  theme_classic()
plot.settings + geom_point(aes(x = height, y = mass, colour = Mainspecies), position = "jitter")
```



e.

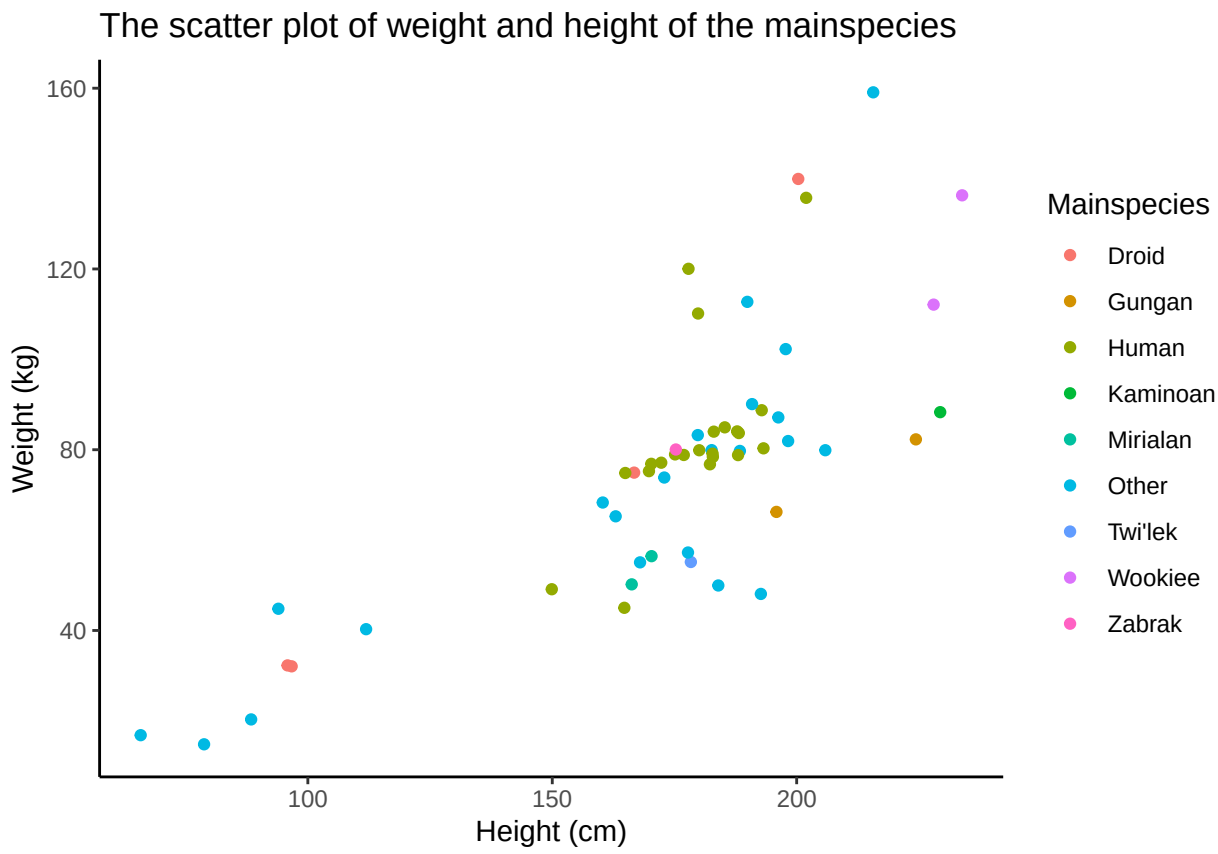
```
outlier <- arrange(starwars, desc(mass)) %>% head(1)

select(outlier, name, height, mass, species)
```

```
## # A tibble: 1 x 4
##   name                  height  mass species
##   <chr>                  <int> <dbl> <chr>
## 1 Jabba Desilijic Tiure    175  1358 Hutt
```

```
plot.settings <- anti_join(starwars,outlier) %>% filter(is.na(Mainspecies)==FALSE & is.na(height)==FALSE
  ggplot(aes(x = height, y = mass)) +
  labs(x = "Height (cm)", y = "Weight (kg)", title = "The scatter plot of weight and height of the mains
  theme_classic()
```

```
## Joining with `by = join_by(name, height, mass, hair_color, skin_color,
## eye_color, birth_year, sex, gender, homeworld, species, films, vehicles,
## starships, num, Mainspecies)`
```

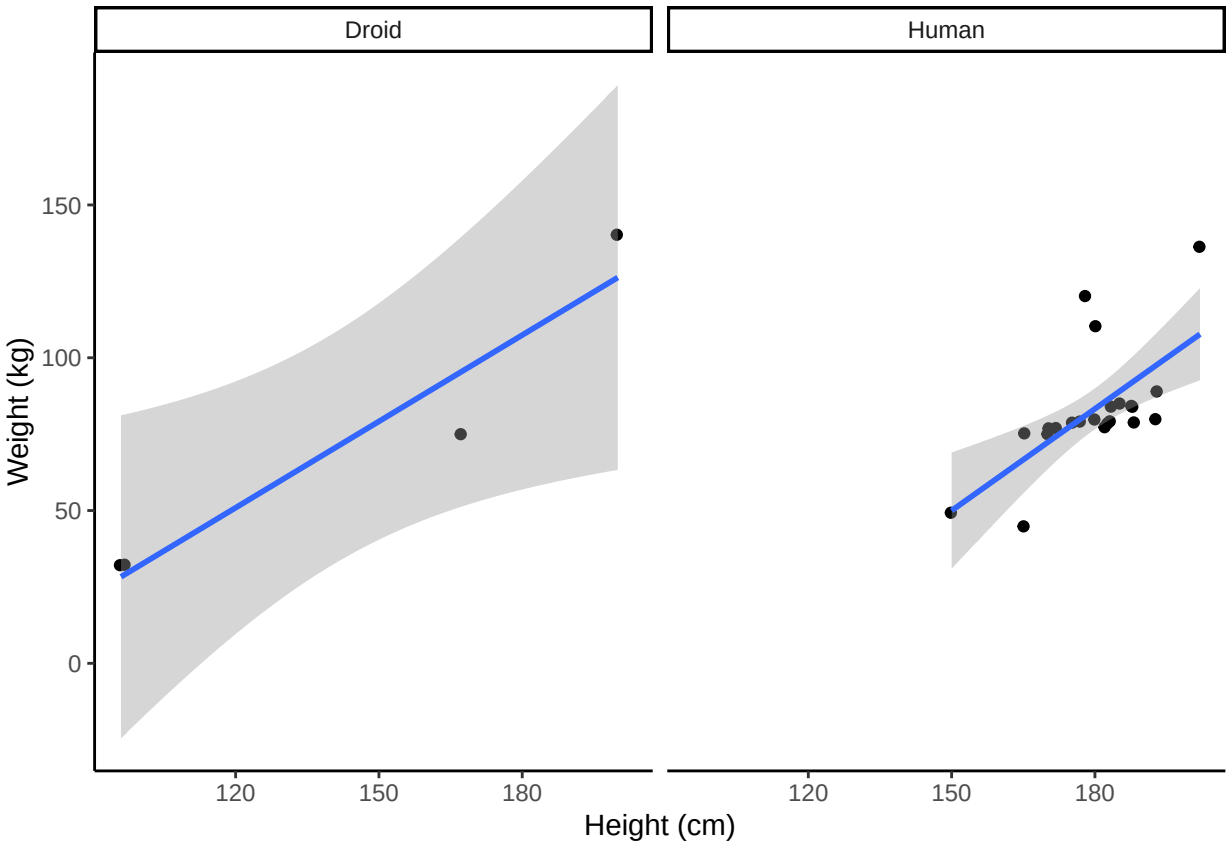
```
plot.settings + geom_point(aes(x = height, y = mass, colour = Mainspecies), position = "jitter")
```



f.

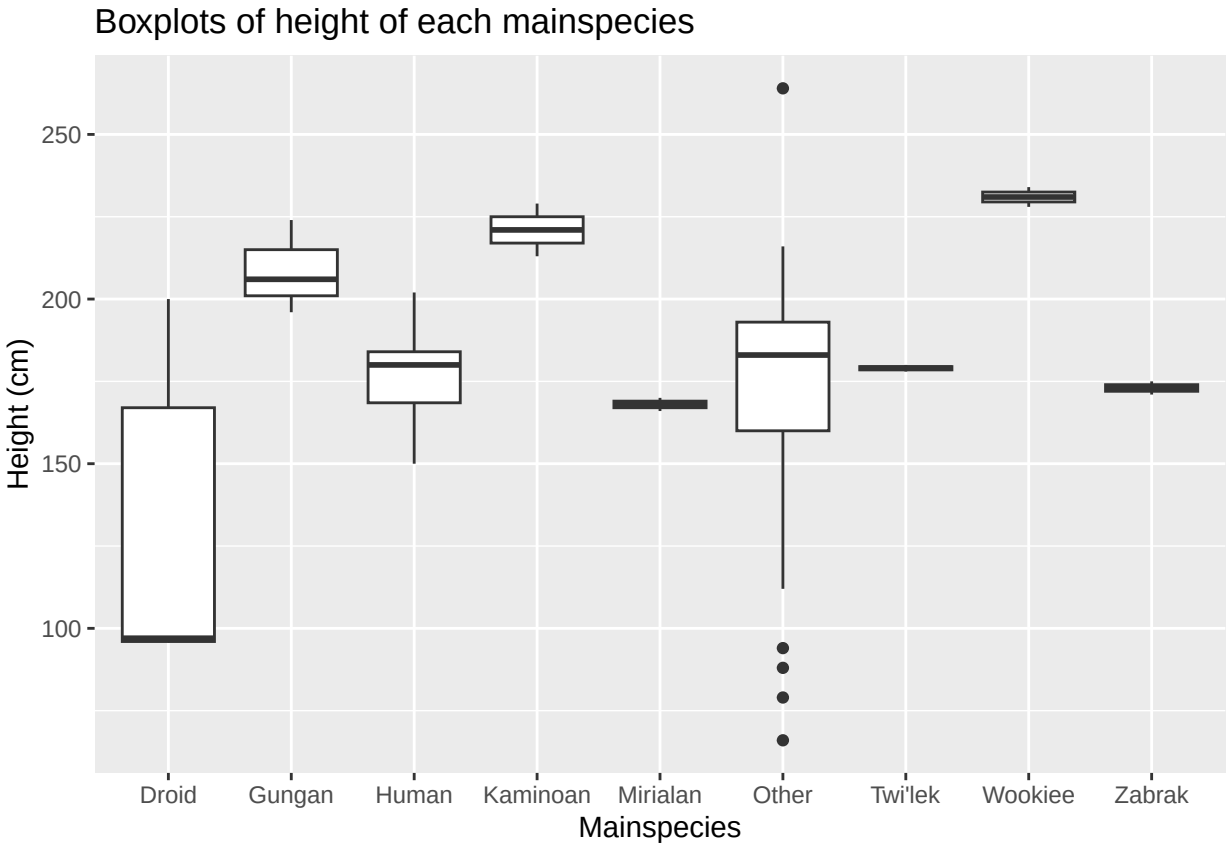
```
plot.settings <- filter(starwars, (species == "Human" | species == "Droid") & is.na(height)==FALSE & is
  labs(x = "Height (cm)", y = "Weight (kg)") +
  theme_classic()
plot.settings +
  geom_point(aes(x = height, y = mass), position = "jitter") +
  facet_wrap(~species) +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



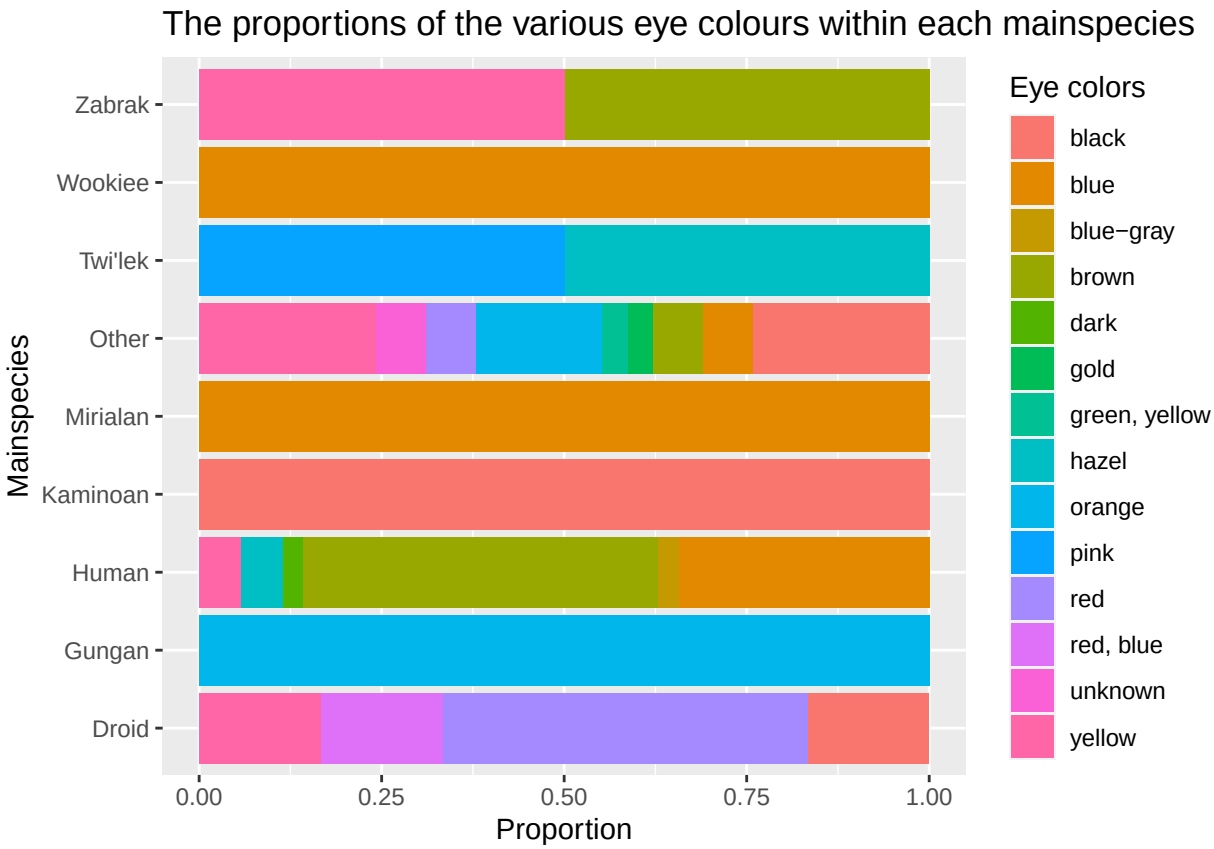
g.

```
filter(starwars, is.na(Mainspecies)==FALSE & is.na(height)==FALSE) %>% ggplot() +
  geom_boxplot(aes(x = Mainspecies, y = height)) +
  labs(x="Mainspecies", y="Height (cm)", title = "Boxplots of height of each mainspecies")
```



h.

```
filter(starwars, is.na(Mainspecies)==FALSE) %>% ggplot() +
  geom_bar(aes(x = Mainspecies, fill = eye_color), position = "fill") +
  coord_flip() +
  labs(x = "Mainspecies", y = "Proportion",fill = "Eye colors", title = "The proportions of the various
```


The proportions of the various eye colours within each mainspecies
Mainspecies
Zabrak
Wookiee
Twi'lek
Other
Mirialan
Kaminoan
Human
Gungan
Droid
0.00
0.25
0.50
0.75
1.00
Proportion
Eye colors
black
blue
blue−gray
brown
dark
gold
green, yellow
hazel
orange
pink
red
red, blue
unknown
yellow