Assignment 1

DABN19: Data Visualization

Dirk Baars

2021-09-30

1 Task 1

First, the data is filtered, whereby only Tatooine homeworld is chosen, via the filter() command. Then, the missing cases are deleted via the drop_na() command, and finally the data is grouped by eye color via grouped_by(), so finally the average mean is calculated via the summarize function per group.

```
starwars_tatooine_summary <-
  starwars %>%
  filter(homeworld == "Tatooine") %>%
  drop_na(mass) %>%
  group_by(eye_color) %>%
  summarize(avg_mass = mean(mass))

ggplot(starwars_tatooine_summary, aes(eye_color, avg_mass)) +
  geom_col() +
  xlab("Eye Color") +
  ylab("Average Weight")
```

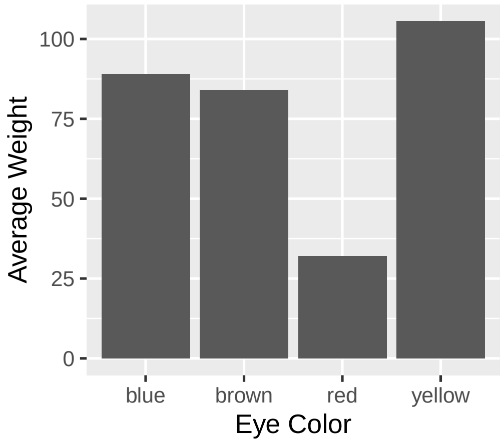

Figure 1: Average weight of Tatooine character by eye color

2 Task 2

2.1 The data in table4a is not tidy, because each variable does not have its own column (year and tuberculosis cases) and each observation does not have its own row. 2.2 In table 4a there are the following variables: Country, year and cases of tuberculosis.

To make the data tidy, it is pivoted in a longer way by using pivot_longer(), whereby the column names are merged into one variable 'year' and the values all got their own row under the column cases.tuberculosis and are separated by years.

```
table4a_tidy <- table4a %>%
  pivot_longer(
    cols = c("1999", "2000"),
    names_to = "year",
    values_to = "cases.tuberculosis"
  )

knitr::kable(table4a_tidy, digits = 5, format.args = list(scientific=
FALSE,big.mark=","),caption =
'Tuberculosis cases in three countries across two years')
```

Table 1: Tuberculosis cases in three countries across two years

country	year	cases.tuberculosis
Afghanistan	1999	745
Afghanistan	2000	2,666
Brazil	1999	37,737
Brazil	2000	80,488
China	1999	212,258
China	2000	213,766

3 Task 3

Finally, table 4b is also made tidy in the same matter as table 4a. Then the two are joined via left_join(). Since the column names are similar, no 'by =' is used. After that a new column is created via mutate(), calculating the rate of tuberculosis.

```
table4b_tidy <- table4b %>%
  pivot_longer(
    cols = c("1999", "2000"),
    names_to = "year",
    values_to = "population"
  )

table4 <- left_join(table4a_tidy, table4b_tidy) %>%
  mutate(rate = cases.tuberculosis/population )

knitr::kable(table4, digits = 5, format.args = list(scientific=FALSE,
big.mark=","),caption = 'Tuberculosis cases in
three countries across two years, including
populations numbers and tuberculosis rate.')
```

Table 2: Tuberculosis cases in three countries across two years, including populations numbers and tuberculosis rate.

country	year	cases.tuberculosis	population	rate
Afghanistan	1999	745	19,987,071	0.00004
Afghanistan	2000	2,666	20,595,360	0.00013
Brazil	1999	37,737	172,006,362	0.00022
Brazil	2000	80,488	174,504,898	0.00046
China	1999	212,258	1,272,915,272	0.00017
China	2000	213,766	1,280,428,583	0.00017