week_3_assignment

RK

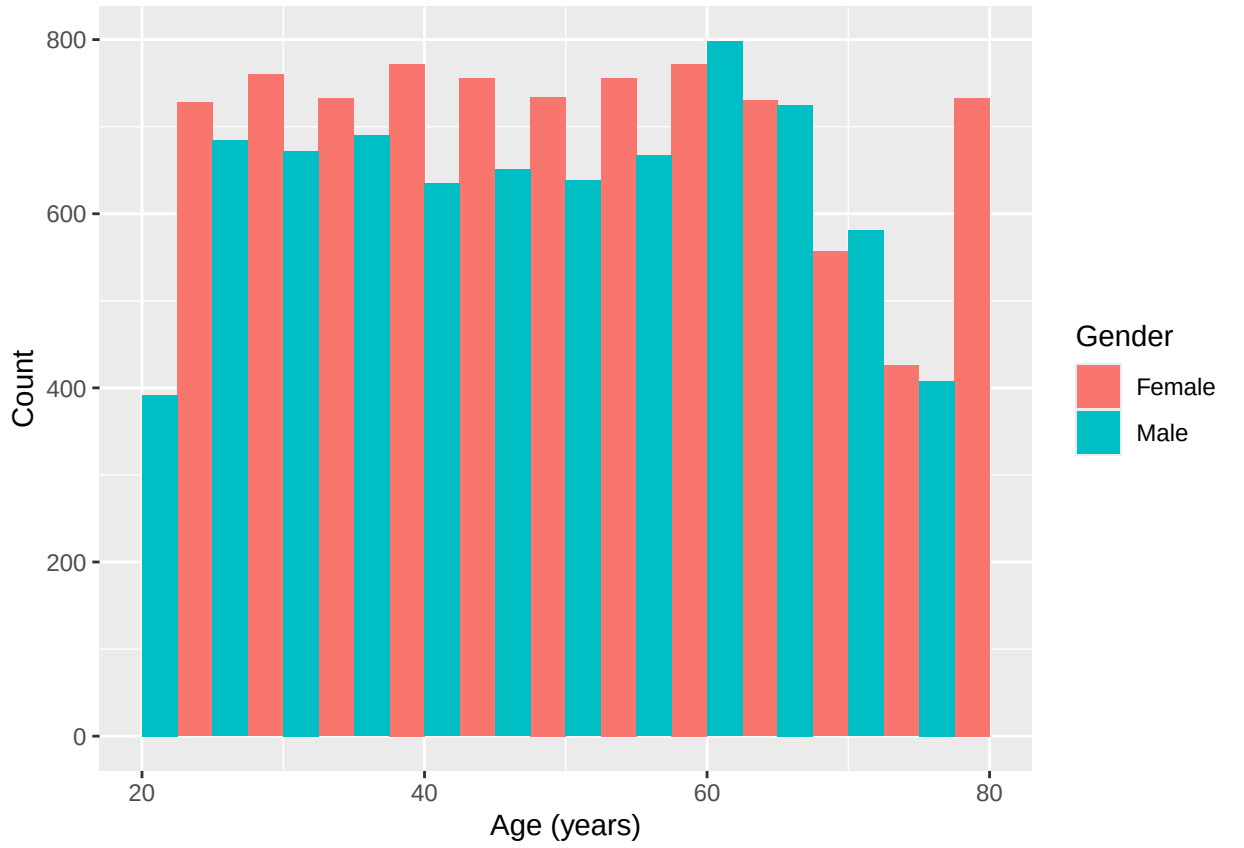
2025-09-21

Data files expected in the working directory: - `cleaned_NHANES.csv` (cleaned NHANES from Week 2) - `diet.csv` (diet study weights over time)

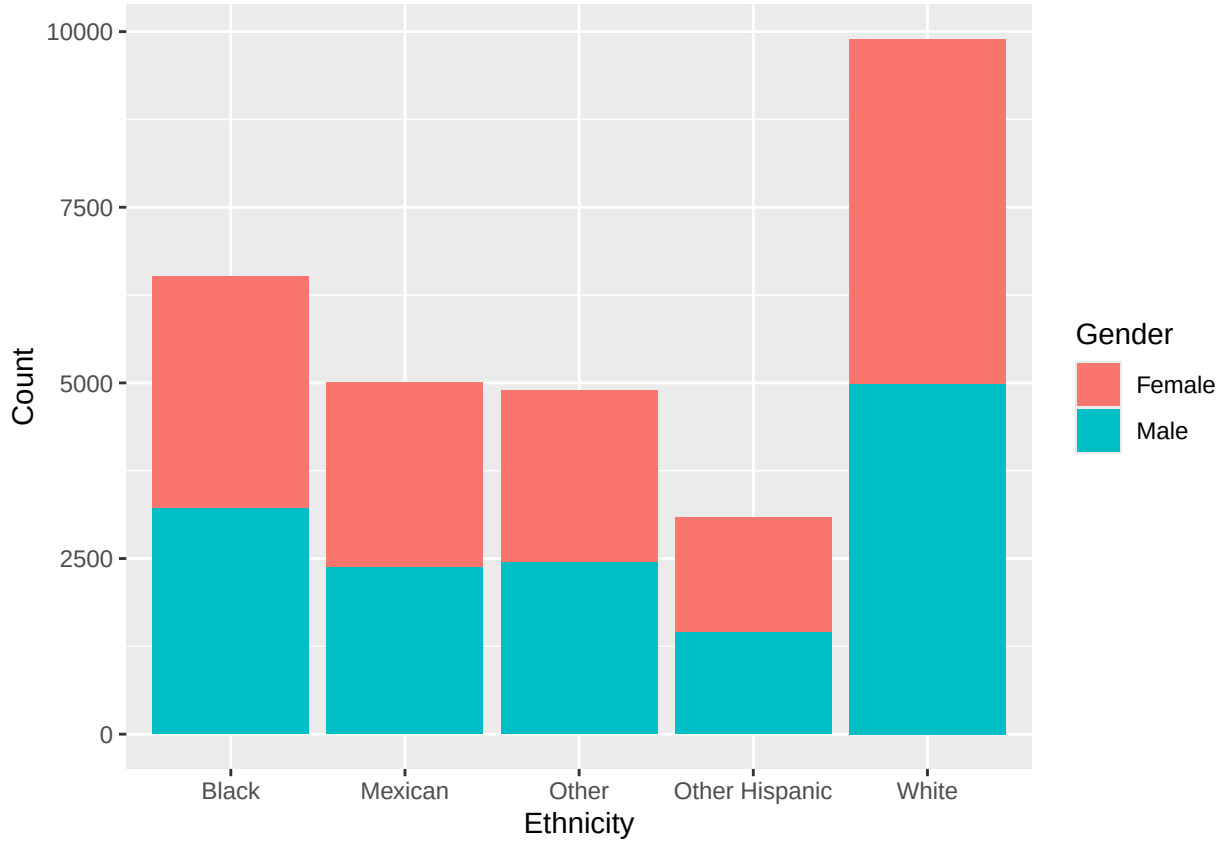
```
# import data sets as provided
nhanes_data <- import(here("assessment", "cleaned_NHANES.csv"))
diet_data <- import(here("assessment", "diet.csv"))
```

Exercise 1 (20%)

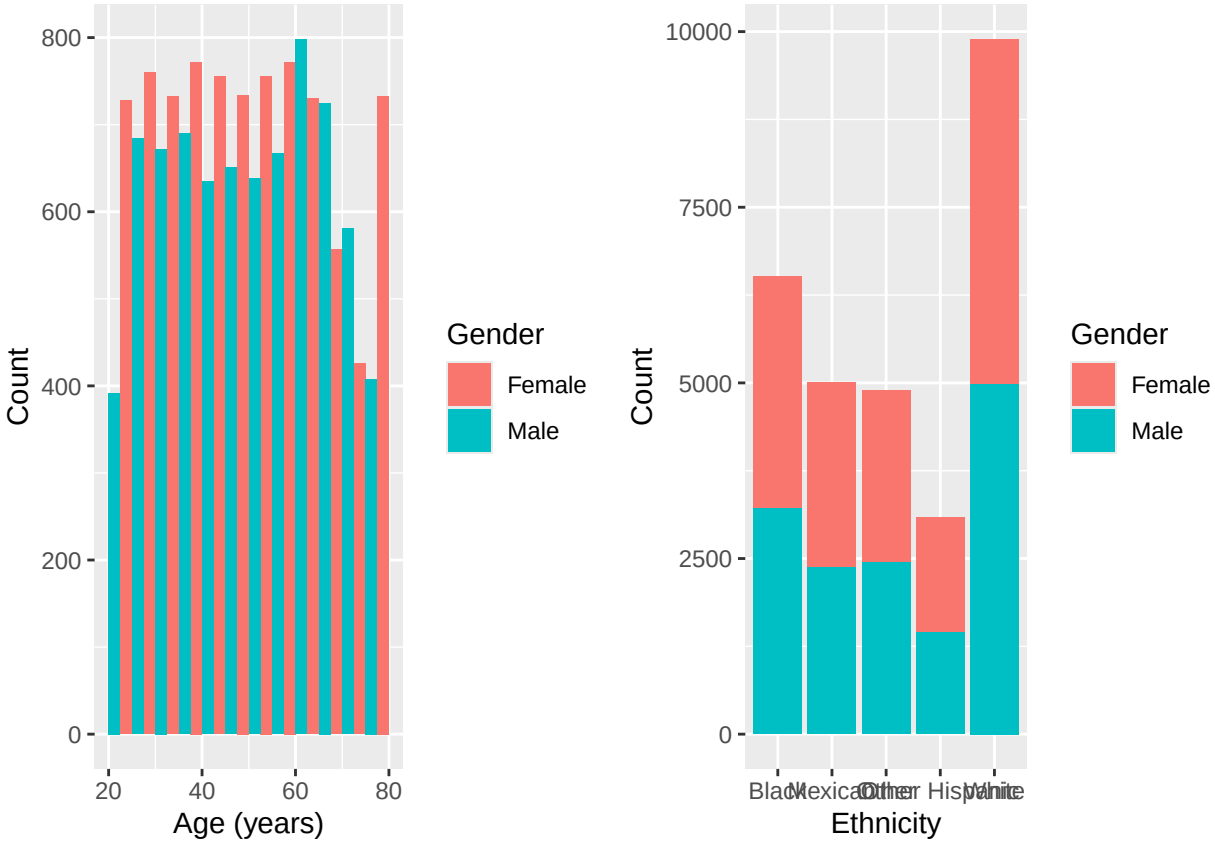
```
## exercise 1

# plot figure 1 (age histogram), group by gender
figure1  <- ggplot(nhanes_data, aes(x = age, fill = gender)) +
  geom_histogram(binwidth = 5, position = "dodge") +
  labs(x = "Age (years)", y = "Count", fill = "Gender") +
  scale_x_continuous(limits = c(20,80), breaks = seq(0, 90, by = 20)) +
  scale_y_continuous(breaks = seq(0,800, by = 200))

# plot figure 2 (ethnicity histogram), group by gender
figure2 <- ggplot(nhanes_data, aes(x = ethnicity_1, fill = gender)) +
  geom_bar() +
  labs(x = "Ethnicity", y = "Count", fill = "Gender")

figure1
```


800
600
400
200
0
Count
20
40
60
80
Age (years)
Gender
Female
Male

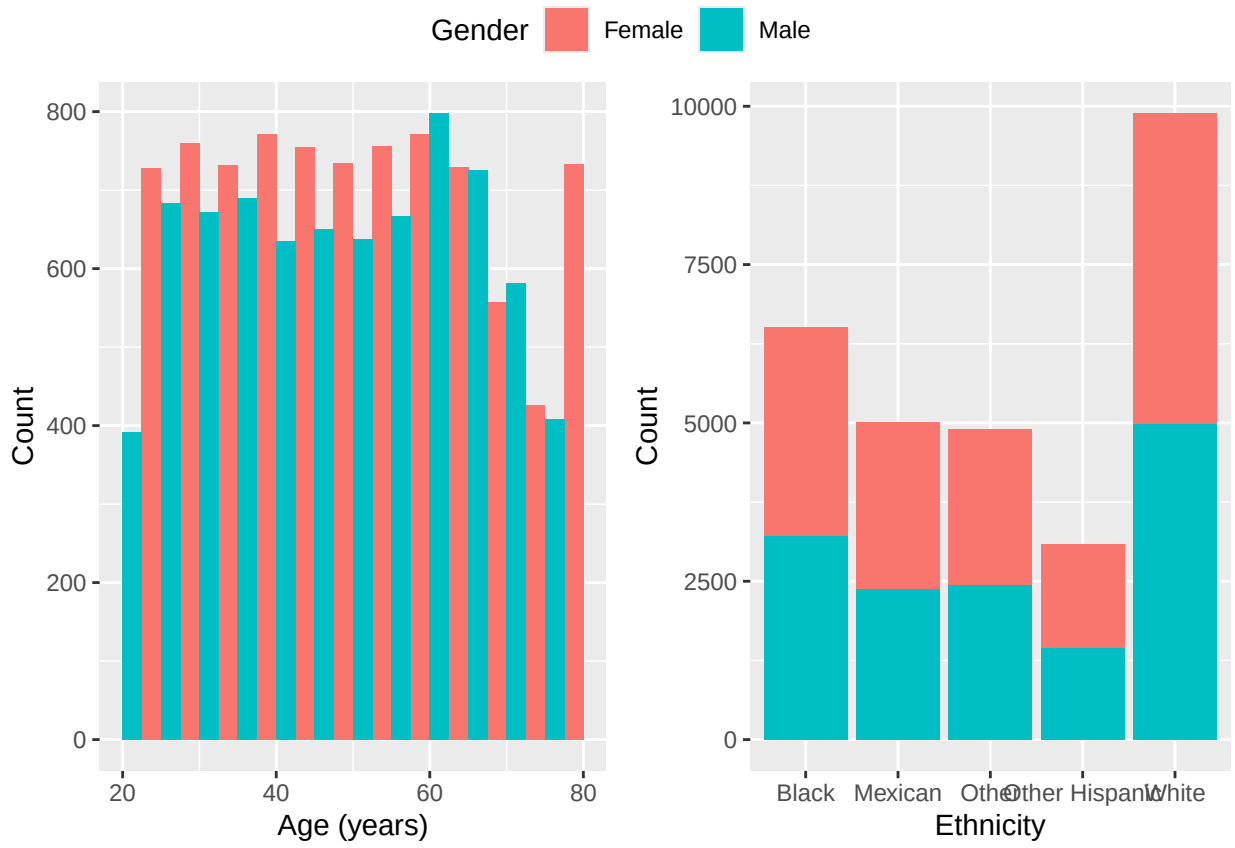
```
figure2
```




```
# combine with cowplot
figure_combined_cow <- cowplot::plot_grid(
  figure1, figure2
  )

figure_combined_cow
```




```
# combine with ggarrange
figure_combined_gg <- ggarrange(
  figure1, figure2,
  common.legend = TRUE, legend = "top"
  )

figure_combined_gg
```



Exercise 1 comparison: The layout of ggarrange looks better than cowplot, because both graphs share a common legend. By combining the legend in the ggarrange output, it makes the output look cleaner and its easier to compare.

Exercise 2 (20%)

```
## age
# create age plot in histogram form
ex2_age <- ggplot(nhanes_data, aes(x = age)) +
  geom_histogram(binwidth = 5, colour = "white", fill = "pink") +
  labs(title = "NHANES Age Distribution", x = "Age (years)", y = "Count")

# determine missing values in age, if any
n_missing(nhanes_data["age"])
```

```
## [1] 0
```

```
## gender
# create gender plot in bar chart form
ex2_gender <- ggplot(nhanes_data, aes(x = gender, fill = gender)) +
  geom_bar() +
  labs(title = "NHANES Gender Distribution", x = "Gender", y = "Count")
```

```r
# determine missing values in gender, if any
n_missing(nhanes_data["gender"])
```

```
## [1] 0
```

```r
## ethnicity

# ethnicity 1 plot
ex2_ethnicity_1 <- ggplot(nhanes_data, aes(x = ethnicity_1)) +
  geom_bar() +
  labs(title = "NHANES Ethnicity Distribution", x = "Ethnicity", y = "Count")

# ethnicity 2 plot
ex2_ethnicity_2 <- ggplot(nhanes_data, aes(x = ethnicity_2)) +
  geom_bar() +
  labs(title = "NHANES Ethnicity Distribution, with Asian category",
       x = "Ethnicity",
       y = "Count")

# determine missing values in ethnicity if any
n_missing(nhanes_data["ethnicity_1"])
```

```
## [1] 0
```

```r
n_missing(nhanes_data["ethnicity_2"])
```

```
## [1] 0
```

Exercise 2 Distributions

Age variable plot: the distribution of age is right-skewed, and there are no missing values. Gender variable plot: the distribution is fairly evenly split between men and women, and there are no missing values. Ethnicity variable plot(s): the distribution for both includes a greater proportion of those whose ethnicity is white, compared to other ethnicities. There are no missing values.

```r
# compare both ethnicity plots

ethnicity_combined <- cowplot::plot_grid(
  ex2_ethnicity_1, ex2_ethnicity_2)


# remove ethnicity_1, see explanation below
nhanes_data <- nhanes_data %>%
  select(-ethnicity_1)
```

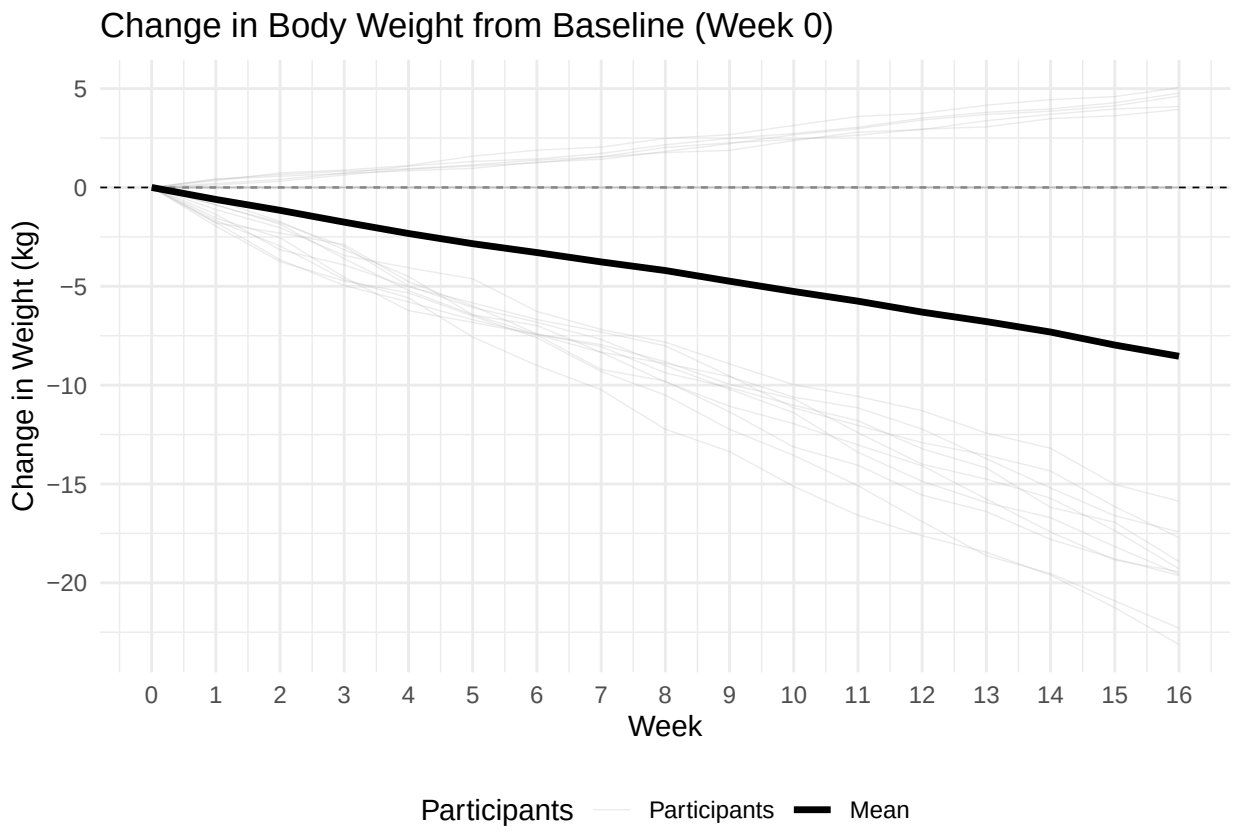
Exercise 2 Ethnicity Comparison

I am keeping the ethnicity_2 variable. It provides greater insight into the demographic specifics of the data. The 'other' category in ethnicity_1 is made up of more than half of those who should be in the 'Asian' category. It is better to have more specifics about the ethnicity of those in the dataset because better category definitions allow for better future analysis.

Exercise 3

Changes to make to current plot: Overplotting/spaghetti: Use light, thin lines per participant and optionally show a thicker mean trend with a smoother/summary to reduce clutter. Also, current graph has differing thickness and colours for participants, that should be uniform. - **Scales & guides:** Clearly label axes (add weight measurement to y axes), add label to participants legend, and add a title for the graph

```
diet_delta <- diet_data %>%
  group_by(Participant) %>%
  mutate(
    baseline_w0 = Weight[Week == 0][1],
    delta_kg    = Weight - baseline_w0
  ) %>%
  ungroup()

# mean change per week (for the bold overlay)
mean_delta <- diet_delta %>%
  group_by(Week) %>%
  summarize(mean_delta = mean(delta_kg, na.rm = TRUE), .groups = "drop")

max_week <- max(diet_delta$Week, na.rm = TRUE)

p_diet <- ggplot(diet_delta, aes(x = Week, y = delta_kg, group = Participant)) +
  geom_hline(yintercept = 0, linetype = "dashed", size = 0.3) +
  # uniform thin lines for all participants; legend entry created via constant mapping
  geom_line(aes(color = "Participants"), size = 0.25, alpha = 0.25) +
  # thicker mean line
  geom_line(
    data = mean_delta,
    aes(x = Week, y = mean_delta, color = "Mean"),
    size = 1.2,
    inherit.aes = FALSE
  ) +
  scale_color_manual(
    name   = "Participants",                    # legend title
    breaks = c("Participants", "Mean"),
    values = c("Participants" = "grey70", "Mean" = "black")
  ) +
  scale_x_continuous(breaks = seq(0, max_week, by = 1)) +
  labs(
    title = "Change in Body Weight from Baseline (Week 0)",
    x     = "Week",
    y     = "Change in Weight (kg)"
  ) +
  theme_minimal(base_size = 11) +
  theme(legend.position = "bottom")

p_diet
```


Change in Body Weight from Baseline (Week 0)
Change in Weight (kg)
5
0
−5
−10
−15
−20
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Week
Participants
Participants
Mean

Exercise 4