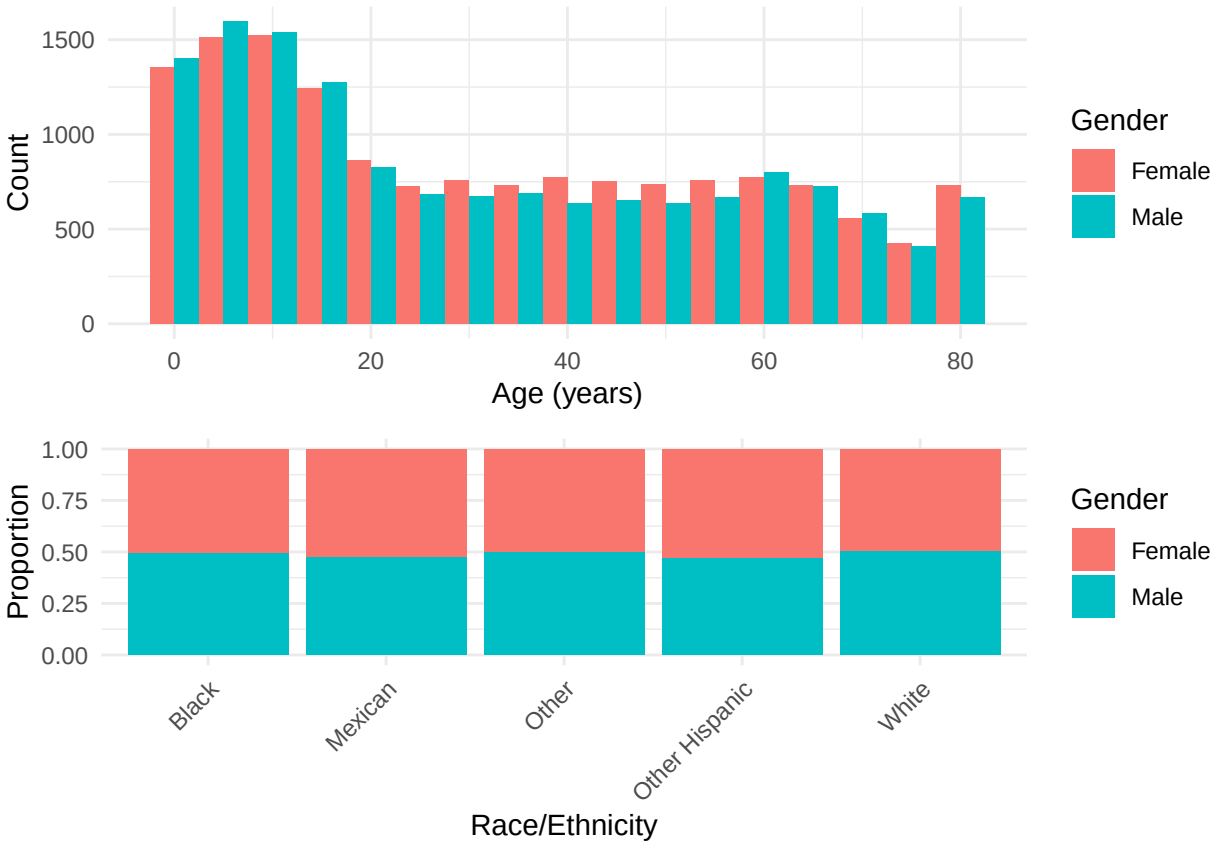
Reproducing Graphs

```
library(dplyr)
library(ggplot2)
library(cowplot)
library(ggpubr)
```

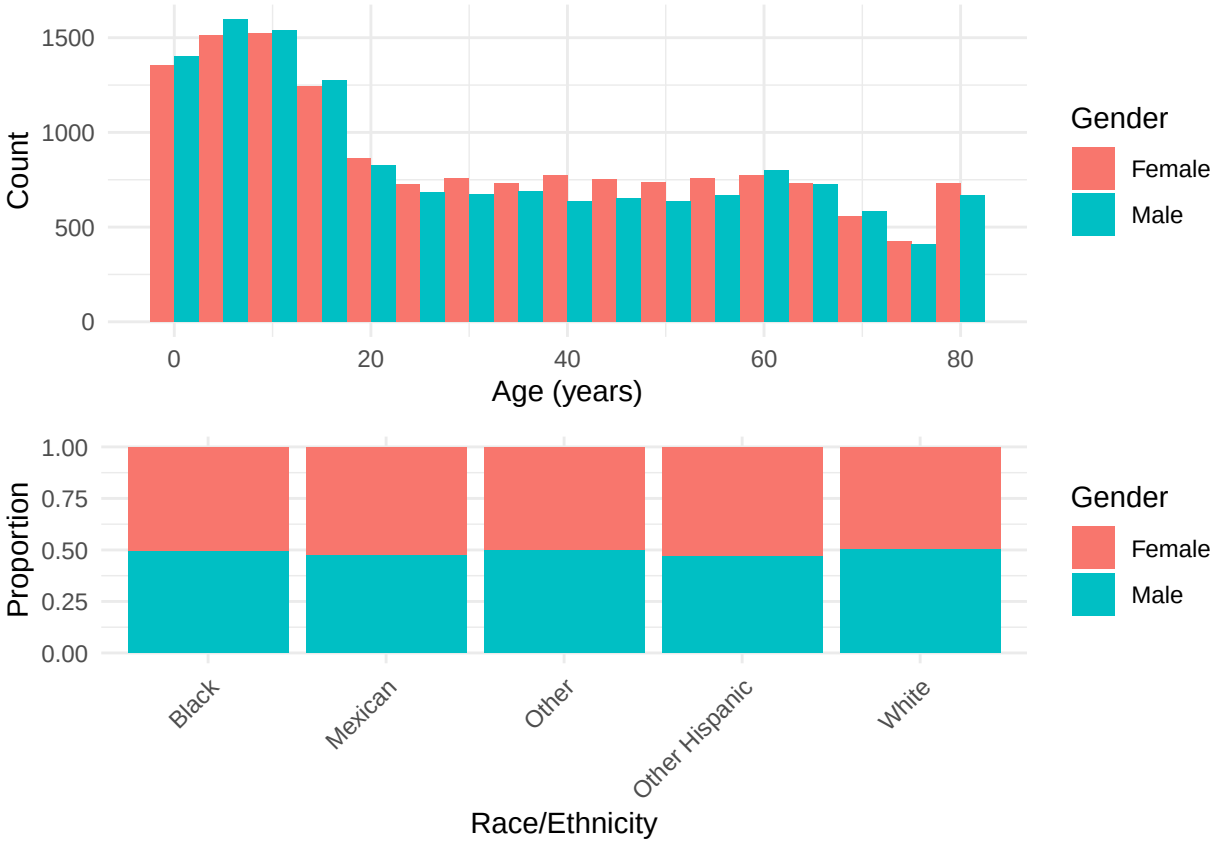
Exercise 1: Reproducing and arranging ggplot2 figures

The two figures attached below shows the plots combined using cowplot as well as ggarrange. Both plots shows exactly same organization of plots showing no difference. Both layouts are same.

```
nhanes_clean <- read.csv("cleaned_NHANES.csv")

# 1st plot: Age distribution by gender (5-year bins)
p1 <- ggplot(nhanes_clean, aes(x = age, fill = gender)) +
  geom_histogram(binwidth = 5, position = "dodge") +
  labs(x = "Age (years)", y = "Count", fill = "Gender") +
  theme_minimal()

# 2nd plot: Ethnicity distribution by gender
p2 <- ggplot(nhanes_clean, aes(x = ethnicity_1, fill = gender)) +
  geom_bar(position = "fill") +
  labs(x = "Race/Ethnicity", y = "Proportion", fill = "Gender") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Combine using cowplot
cow_combined <- plot_grid(p1, p2, ncol = 1)

# Combine using ggpubr
ggpubr_combined <- ggarrange(p1, p2, ncol = 1, nrow = 2)

# Show results
print(cow_combined)
```




```
print(ggpubr_combined)
```



Exercise # 02):

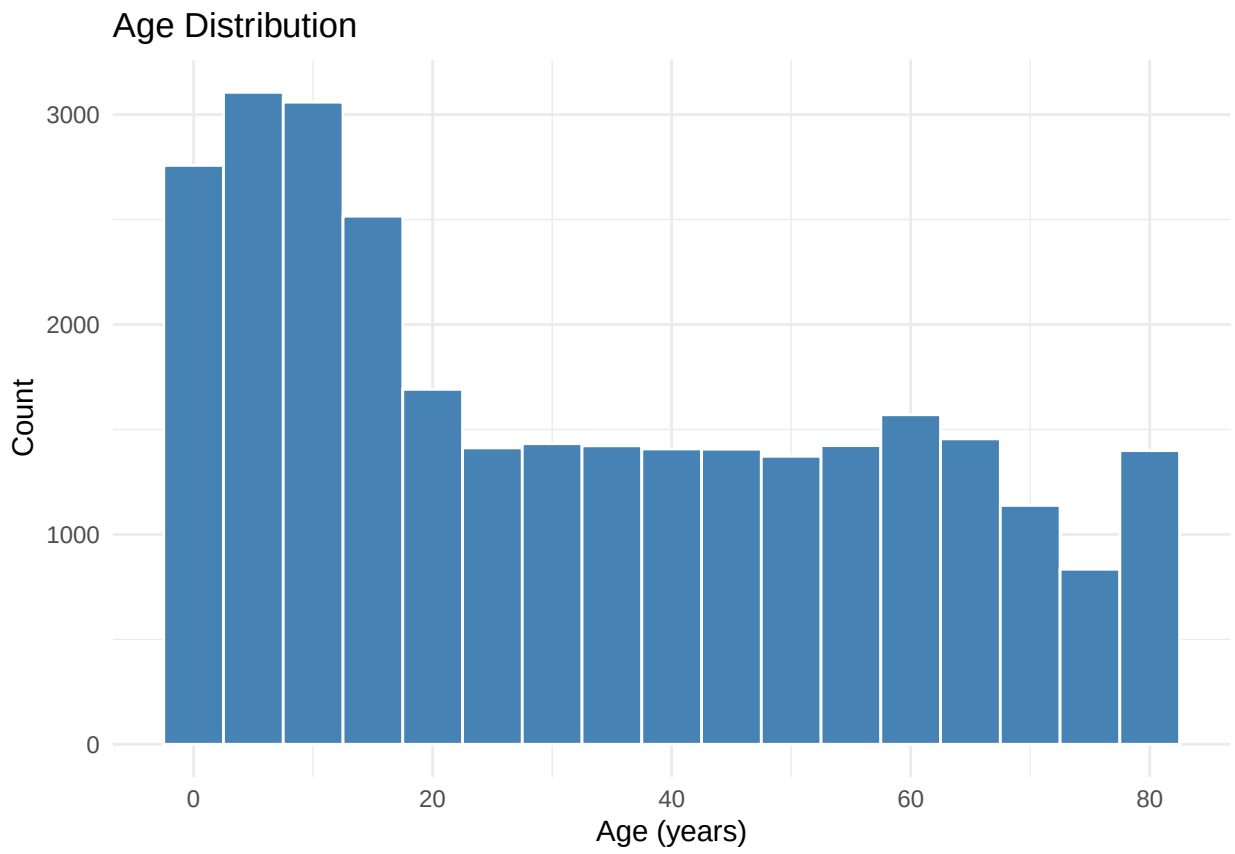
There are no missing values in any four columns of the dataset.

```
colSums(is.na(nhanes_clean)[, c("age", "gender", "ethnicity_1", "ethnicity_2")])
```

```
##         age      gender ethnicity_1 ethnicity_2
##           0           0           0           0
```

The histogram attached below for age shows that the distribution is positively skewed with a maximum range of 80. Most of the people have age below 20.

```
p_age <- ggplot(nhanes_clean, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Age Distribution",
       x = "Age (years)",
       y = "Count") +
  theme_minimal()

ggsave("age_distribution.png", p_age, width = 7, height = 5)

p_age
```

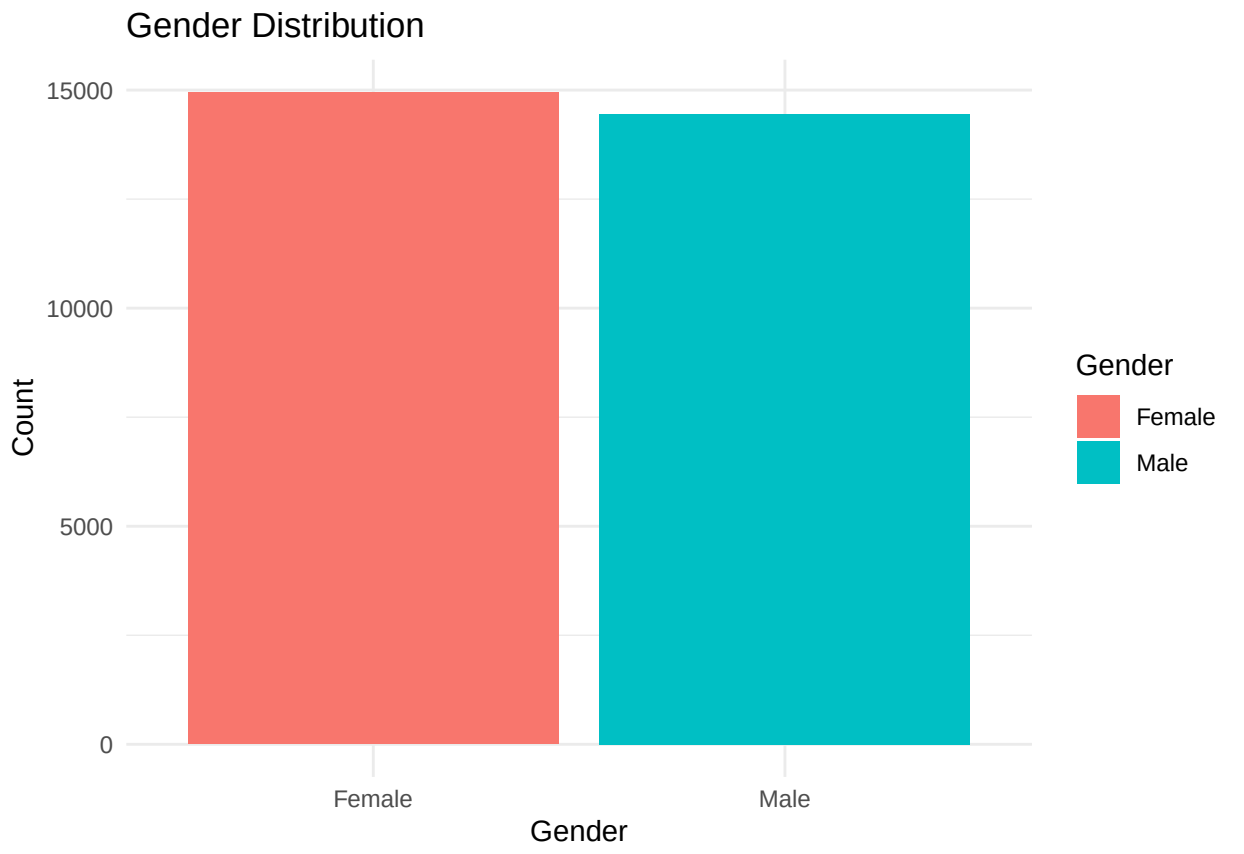



```
# Summary
summary(nhanes_clean$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   10.00   28.00   32.52   54.00   80.00
```

The frequency distribution plot for gender shows that the frequency of female is slightly higher compared to male.

```
p_gender <- ggplot(nhanes_clean, aes(x = gender, fill = gender)) +
  geom_bar() +
  labs(title = "Gender Distribution",
       x = "Gender",
       y = "Count",
       fill = "Gender") +
  theme_minimal()

p_gender
```

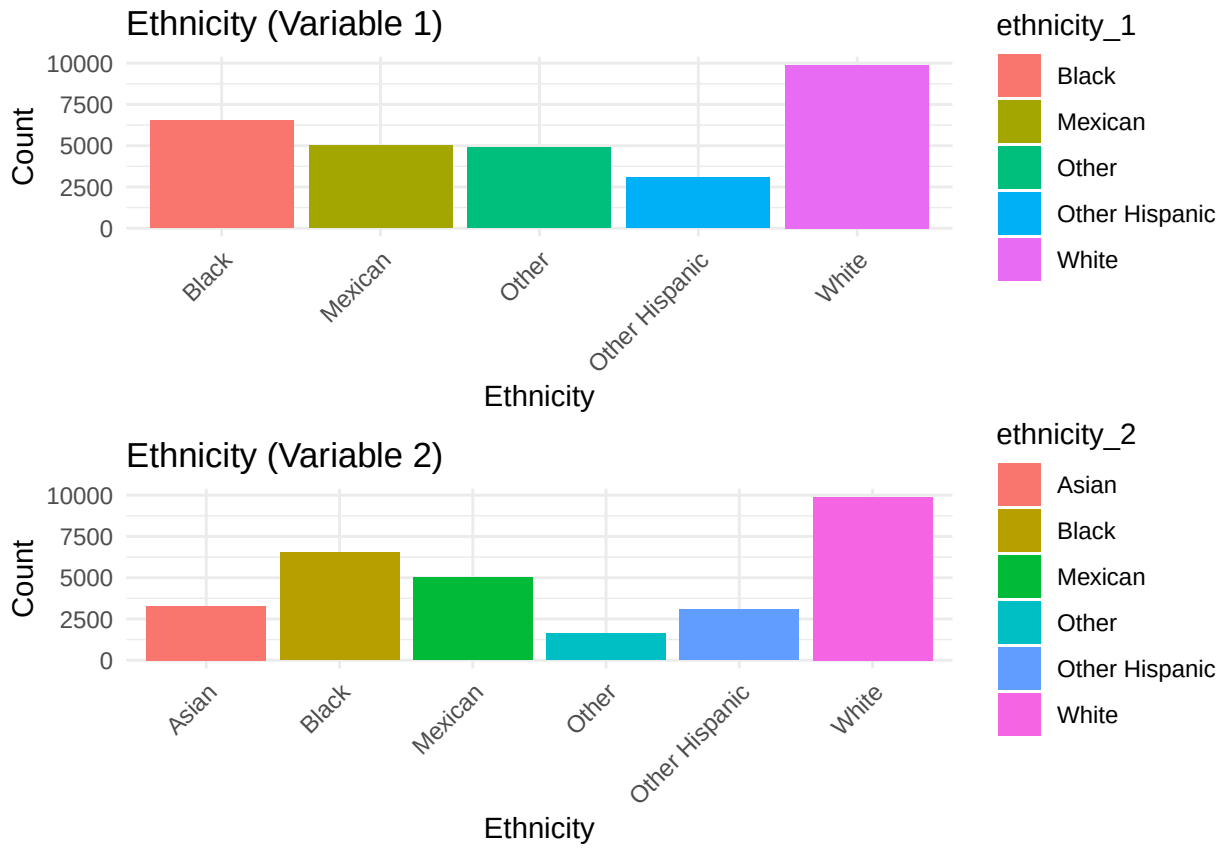



```
ggsave("gender_distribution.png", p_gender, width = 6, height = 4)
```

Ethnicity 1 included the categories Black, Mexican, Other, Other Hispanic, and White. Ethnicity 2, on the other hand, provided the categories Asian, Black, Mexican, Other, Other Hispanic, and White. While both variables covered similar groups, Ethnicity 2 offered greater specificity by explicitly distinguishing Asian participants rather than grouping them under a broad "Other" category. This disaggregation improves interpretability and aligns more closely with standard NHANES reporting conventions, which typically separate Asian individuals into their own category. For these reasons, Ethnicity 2 was retained for further analyses, and Ethnicity 1 was discarded.

```
# Ethnicity 1
p_eth1 <- ggplot(nhanes_clean, aes(x = ethnicity_1, fill = ethnicity_1)) +
  geom_bar() +
  labs(title = "Ethnicity (Variable 1)",
       x = "Ethnicity",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

ggsave("ethnicity_1_distribution.png", p_eth1, width = 8, height = 5)

# Ethnicity 2
p_eth2 <- ggplot(nhanes_clean, aes(x = ethnicity_2, fill = ethnicity_2)) +
  geom_bar() +
  labs(title = "Ethnicity (Variable 2)",
```

```
       x = "Ethnicity",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

plot_grid(p_eth1, p_eth2, ncol = 1)
```




```
ggsave("ethnicity_2_distribution.png", p_eth2, width = 8, height = 5)
```

Exercise # 03):

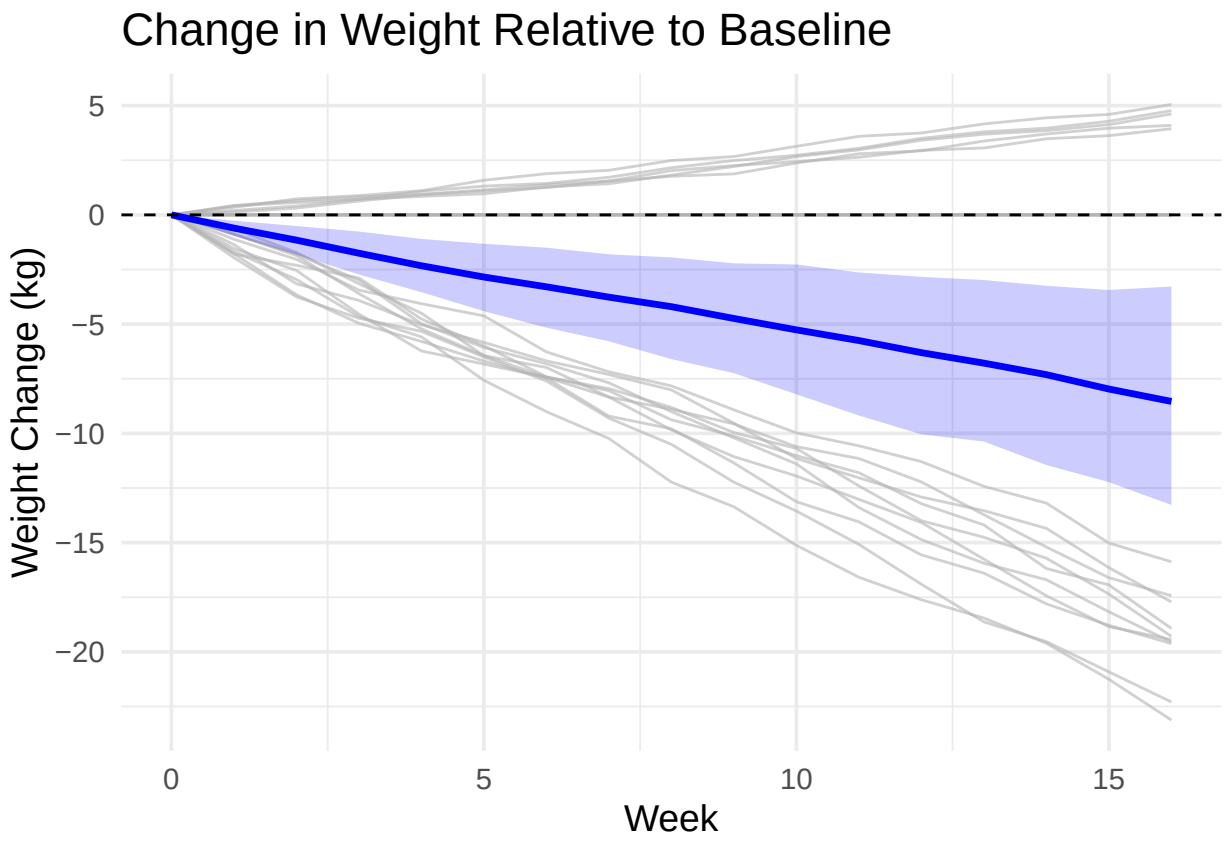
Problems with the Current Figure are:

- Too many overlapping colors: Each participant is given a unique color, which creates a very busy legend and makes it nearly impossible to distinguish lines. With 20 participants, many colors look similar, reducing clarity.
- No baseline reference: Since the research question is about change relative to baseline (week 0), the plot should highlight how weight changes from that starting point. Right now, we only see raw weights.
- Overcrowded legend: The legend includes all 20 participants, which is not reader-friendly and takes up space without adding value. Hard to follow individual trends. Thin lines in similar colors make it difficult to track a single participant across weeks.

Improvements I Would Make are:

- Plot weight change relative to baseline: Subtract each participant's weight at week 0 from all subsequent measurements. This way, the y-axis directly shows loss/gain from baseline.
- Use fewer, clearer colors: Instead of unique colors per participant, make all participants grey lines for context and highlight the average trend (e.g., bold colored line with confidence interval).
- Add a reference line at 0: A horizontal line at 0 makes it easy to see who gained or lost weight.
- Improve readability: Reduce or remove the participant legend (not useful here). Increase line thickness for the average trend.

```r
diet <- read.csv("diet.csv")

# Calculate weight change from baseline
diet <- diet %>%
  group_by(Participant) %>%
  mutate(baseline = Weight[Week == 0],
         weight_change = Weight - baseline)

# Plot
p <- ggplot(diet, aes(x = Week, y = weight_change, group = Participant)) +
  geom_line(color = "grey70", alpha = 0.6) +  # individual trajectories
  stat_summary(fun = mean, geom = "line", aes(group = 1),
               color = "blue", size = 1.2) +  # average trend
  stat_summary(fun.data = mean_cl_boot, geom = "ribbon",
               aes(group = 1), fill = "blue", alpha = 0.2) + # CI ribbon
  geom_hline(yintercept = 0, linetype = "dashed", color = "black") +
  labs(title = "Change in Weight Relative to Baseline",
       x = "Week",
       y = "Weight Change (kg)") +
  theme_minimal(base_size = 14)
p
```




```
ggsave("improved_diet_plot.png", p, width = 8, height = 6)
```

Report: Exploring Relationships Between Age, Gender, and Systolic Blood Pressure

Blood pressure is an important indicator of cardiovascular health, and both age and gender are known to influence risk levels. In this analysis, I examined systolic blood pressure (SBP) readings in the NHANES dataset, focusing on how average SBP varies across age groups and genders, and how individuals are distributed across clinical hypertension categories.

Constructing the Main Variable of Interest

Each participant had up to four SBP readings. To obtain a more reliable estimate, I calculated the average SBP for each person. This measure was set to missing only when all four readings were unavailable. After filtering to include only individuals with non-missing values for age, gender, and average SBP, the sample size decreased compared to the full dataset. Participants were classified into four clinically meaningful groups:

- Normal: SBP $<$ 120 mm Hg
- Elevated: 120–129 mm Hg
- Stage 1 Hypertension: 130–139 mm Hg
- Stage 2 Hypertension: $>=$ 140 mm Hg

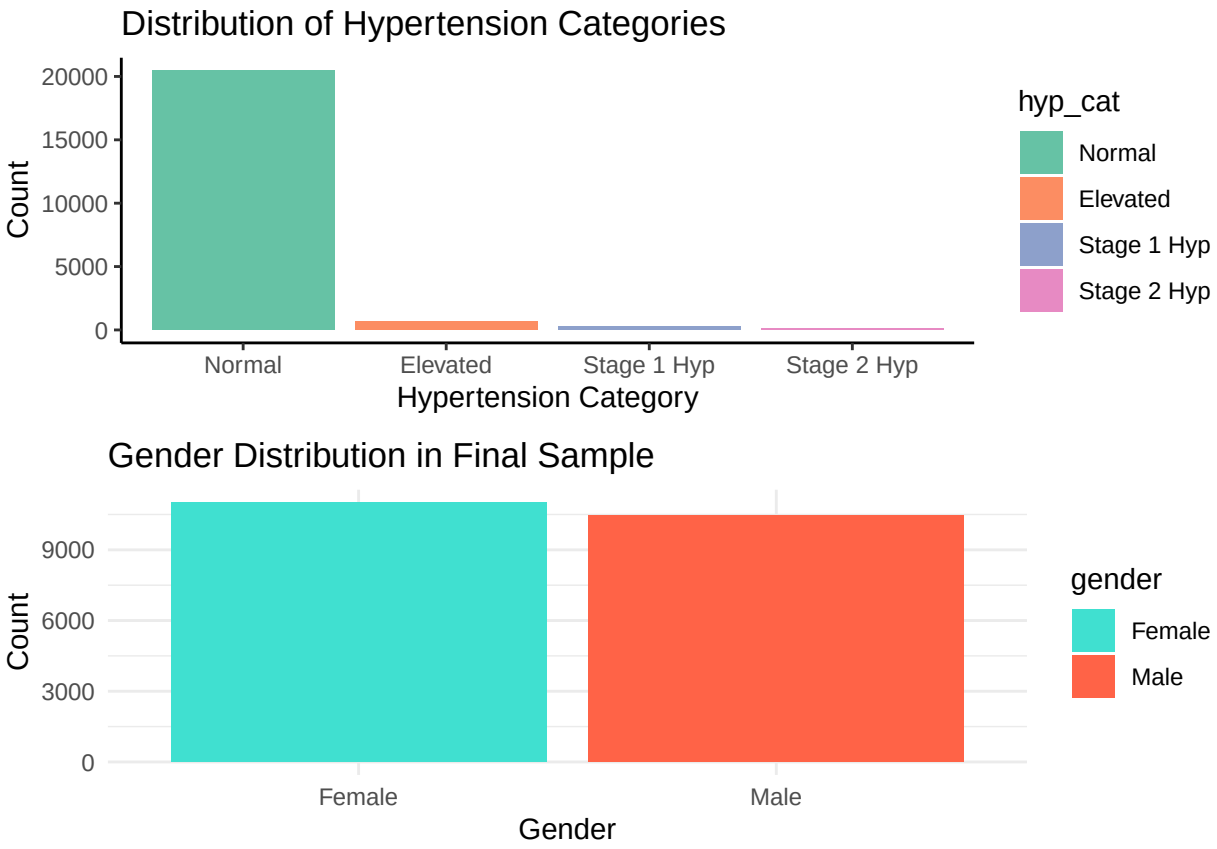
These categories allow us to connect SBP to health risk levels in a straightforward way.

```
nhanes_clean <- nhanes_clean %>%
  mutate(avg_sbp = rowMeans(select(., systolic_bp_1:systolic_bp_4), na.rm = TRUE),
         avg_sbp = ifelse(rowSums(is.na(select(., systolic_bp_1:systolic_bp_4))) == 4, NA,
                          avg_sbp)) %>%
  filter(!is.na(age), !is.na(gender), !is.na(avg_sbp)) %>%
  mutate(hyp_cat = case_when(
    avg_sbp < 120 ~ "Normal",
    avg_sbp >= 120 & avg_sbp <= 129 ~ "Elevated",
    avg_sbp >= 130 & avg_sbp <= 139 ~ "Stage 1 Hyp",
    avg_sbp >= 140 ~ "Stage 2 Hyp"
  ),
  hyp_cat = factor(hyp_cat, levels = c("Normal", "Elevated",
                                       "Stage 1 Hyp",
                                       "Stage 2 Hyp")))
```

Describing the Final Sample

To better understand the cleaned sample, I visualized the distribution of hypertension categories and gender. Normal and elevated blood pressure categories were more common, while a very small portion of the population fell into Stage 1 or Stage 2 hypertension, highlighting the low burden of high blood pressure. The gender distribution was roughly balanced, with only a small change compared to the full dataset.

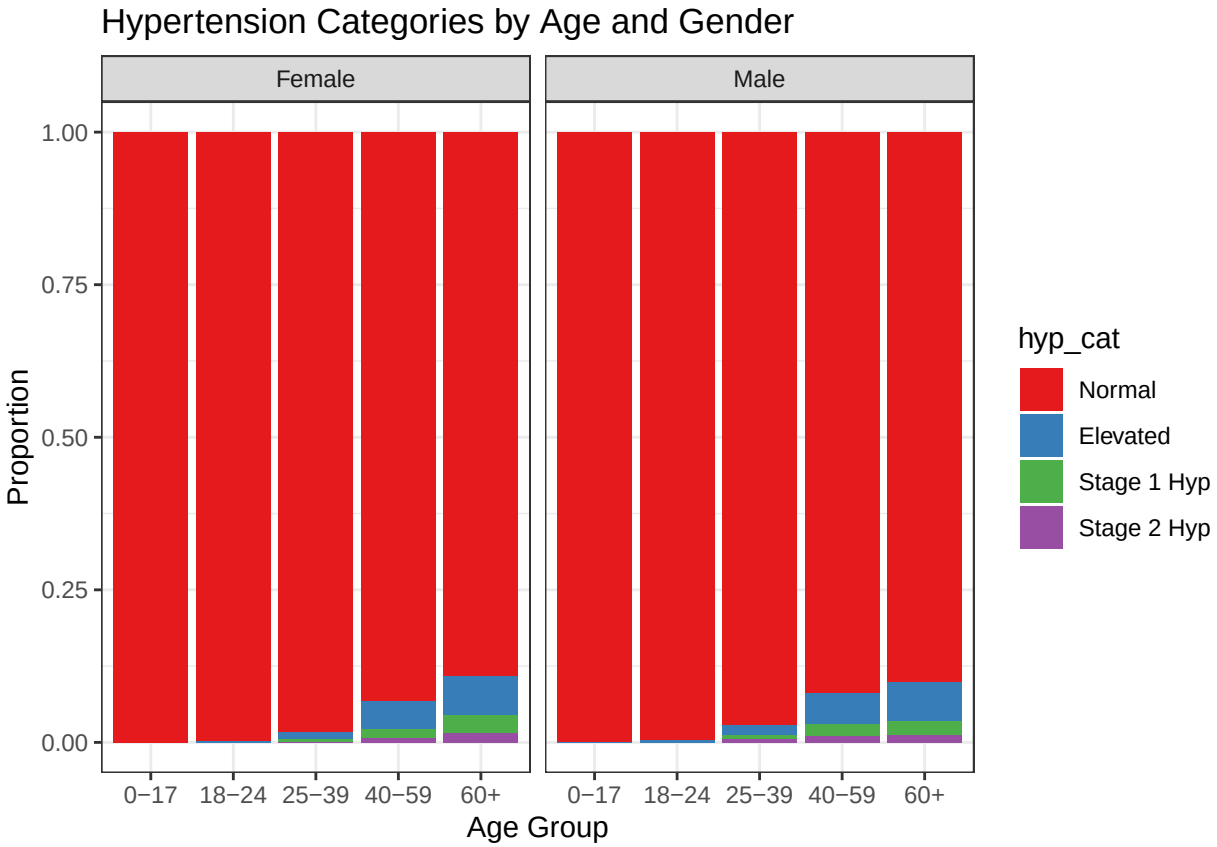
```
nhanes_clean <- nhanes_clean %>%
  select(hyp_cat, age_cat, avg_sbp, gender)

nhanes_clean <- na.omit(nhanes_clean)
# Hypertension category counts
p1 <- ggplot(nhanes_clean, aes(x = hyp_cat, fill = hyp_cat)) +
  geom_bar() +
  labs(title = "Distribution of Hypertension Categories",
       x = "Hypertension Category", y = "Count") +
  scale_fill_brewer(palette = "Set2") +
  theme_classic()

# Gender distribution
p2 <- ggplot(nhanes_clean, aes(x = gender, fill = gender)) +
  geom_bar() +
  labs(title = "Gender Distribution in Final Sample",
       x = "Gender", y = "Count") +
  scale_fill_manual(values = c("Male" = "tomato", "Female" = "turquoise")) +
  theme_minimal()


plot_grid(p1, p2, ncol = 1)
```



Hypertension Across Age and Gender

The graph shows the distribution of hypertension categories by age group and gender. Across both males and females, the majority of individuals fall into the "Normal" category, particularly in the younger age groups (0–39 years), where hypertension prevalence is minimal. However, beginning from the 40–59 age group and more prominently in the 60+ group, there is a noticeable increase in elevated blood pressure and hypertension stages (Stage 1 and Stage 2), indicating age as a strong risk factor. Both genders display similar trends, though the proportions of elevated and hypertensive cases are slightly more visible in older males compared to females. Overall, the chart highlights that hypertension is relatively rare at younger ages but becomes more common in older adults, regardless of gender.

```
p3 <- ggplot(nhanes_clean, aes(x = age_cat, fill = hyp_cat)) +
  geom_bar(position = "fill") +
  facet_wrap(~ gender) +
  labs(title = "Hypertension Categories by Age and Gender",
       x = "Age Group", y = "Proportion") +
  scale_fill_brewer(palette = "Set1") +
  theme_bw()

p3
```



Conclusion

By cleaning and categorizing SBP data, we were able to see clear patterns in the distribution of hypertension across age and gender. Most notably, blood pressure increases steadily with age, leading to a shift toward more severe hypertension categories in older adults. Gender differences are modest but suggest different trajectories across the lifespan. Together, these findings highlight the value of stratified visualizations in communicating health disparities and risks in an accessible way.