T test and Box plot and Life Expectancy Gap

Riley

```r
# Load necessary libraries
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(ggrepel)  # install if necessary
```

```r
# Load CSV
df <- read.csv("/Users/rileynewton/Desktop/cleandata.csv")  # update path

# Remove unwanted columns
df <- df %>% select(-c(X, X.1, X.2, X.3, X.4, X.5, X.6))

# Create life expectancy gap
df$LifeExpectancyGap <- df$`Female.Life.Expectancy..years.` - df$`Male.Life.Expectancy..years
```

```r
# Paired t-test: do women live longer than men?
t_test_result <- t.test(
  df$`Female.Life.Expectancy..years.`,
  df$`Male.Life.Expectancy..years.`,
```

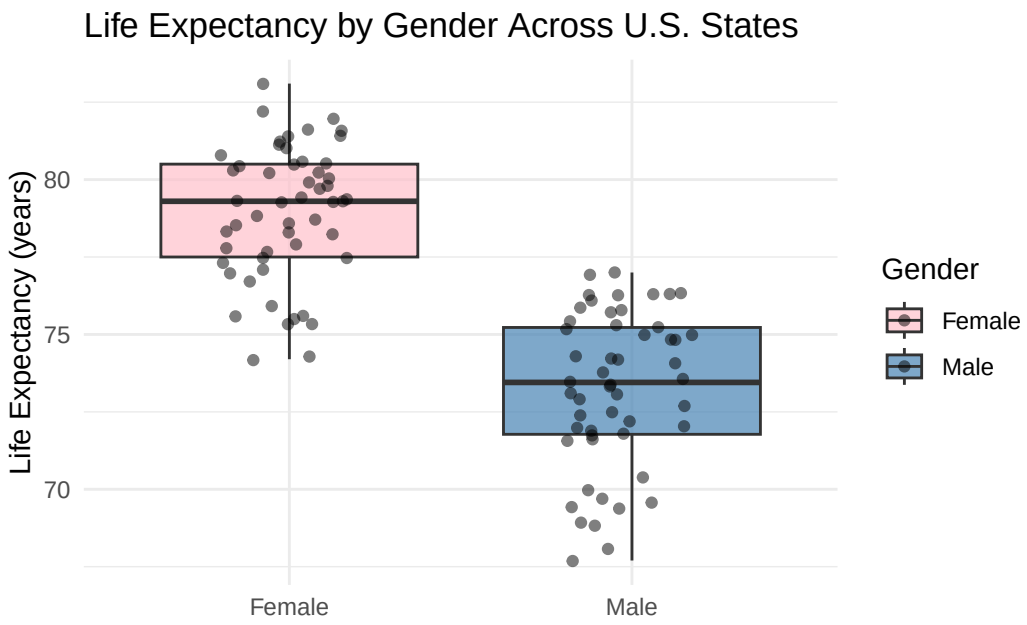
```r
  paired = TRUE,
  alternative = "greater"
)
t_test_result
```

```
    Paired t-test

data:  df$Female.Life.Expectancy..years. and df$Male.Life.Expectancy..years.
t = 68.011, df = 51, p-value < 2.2e-16
alternative hypothesis: true mean difference is greater than 0
95 percent confidence interval:
 5.544588      Inf
sample estimates:
mean difference
       5.684615
```

```r
# Boxplot of male vs female life expectancy
df_long <- df %>%
  select(Area, `Male.Life.Expectancy..years.`, `Female.Life.Expectancy..years.`) %>%
  tidyr::pivot_longer(cols = c(`Male.Life.Expectancy..years.`, `Female.Life.Expectancy..years
                      names_to = "Gender", values_to = "LifeExpectancy") %>%
  mutate(Gender = ifelse(Gender == "Male.Life.Expectancy..years.", "Male", "Female"))

ggplot(df_long, aes(x = Gender, y = LifeExpectancy, fill = Gender)) +
  geom_boxplot(alpha = 0.7) +
  geom_jitter(width = 0.2, alpha = 0.5) +
  labs(title = "Life Expectancy by Gender Across U.S. States",
       y = "Life Expectancy (years)",
       x = "") +
  theme_minimal() +
  scale_fill_manual(values = c("Male" = "steelblue", "Female" = "pink"))
```



A paired-samples t-test showed that women have a significantly higher life expectancy than men across U.S. states, $\mathbf{t}(51) = 68.01$, $\mathbf{p} < .001$. On average, women live **5.68 years longer** than men (95% CI: 5.54, ∞).

Finding if Female-Male Age gaps is larger in states with lower total life expectancy:

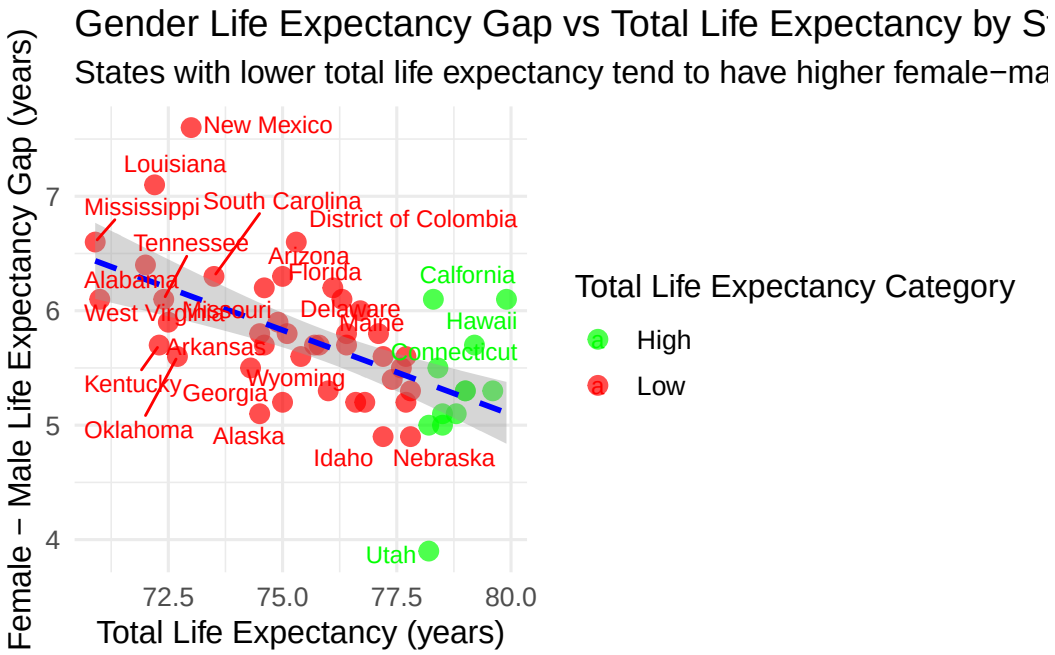
```
# Advanced life expectancy gap plot
df <- df %>%
  mutate(LifeExpCategory = ifelse(`Total.Life.Expectancy..years.` < 78, "Low", "High"))

gap_plot <- ggplot(df, aes(x = `Total.Life.Expectancy..years.`, y = LifeExpectancyGap, color
  geom_point(size = 3, alpha = 0.7) +
  geom_smooth(method = "lm", se = TRUE, color = "blue", linetype = "dashed") +
  geom_text_repel(aes(label = Area), size = 3, max.overlaps = 15) +
  scale_color_manual(values = c("Low" = "red", "High" = "green")) +
  labs(
    title = "Gender Life Expectancy Gap vs Total Life Expectancy by State",
    subtitle = "States with lower total life expectancy tend to have higher female-male gaps"
    x = "Total Life Expectancy (years)",
    y = "Female - Male Life Expectancy Gap (years)",
    color = "Total Life Expectancy Category"
  ) +
  theme_minimal(base_size = 12)
```

```
print(gap_plot)
```

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

```
Warning: ggrepel: 27 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



- cor = -0.577 This is a moderate-to-strong negative correlation.
  - Interpretation: as total life expectancy decreases, the female–male gap tends to increase. In other words, states where people live shorter lives tend to have a bigger difference between women's and men's life expectancy.
- t = -4.99, df = 50, p-value = 7.688e-06
  - The p-value is extremely small (< 0.001).
  - This means the correlation is statistically significant — it's very unlikely this negative relationship is due to chance.
- -0.734 to -0.360

- This tells you that, with 95% confidence, the true correlation in the population is between -0.73 and -0.36.
- Since the entire interval is negative, it confirms the negative relationship.

```r
# Linear regression: LifeExpectancyGap ~ Total.Life.Expectancy
lm_model <- lm(LifeExpectancyGap ~ `Total.Life.Expectancy..years.`, data = df)
summary(lm_model)
```

```
Call:
lm(formula = LifeExpectancyGap ~ Total.Life.Expectancy..years.,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-1.45894 -0.31628 -0.01462  0.19358  1.47396

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   16.89489    2.24750   7.517 9.34e-10 ***
Total.Life.Expectancy..years. -0.14752    0.02956  -4.990 7.69e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4974 on 50 degrees of freedom
Multiple R-squared:  0.3325,    Adjusted R-squared:  0.3191
F-statistic:  24.9 on 1 and 50 DF,  p-value: 7.688e-06
```