STAT 300: Written Assignment 2

Aronn Grant Laurel (21232475)

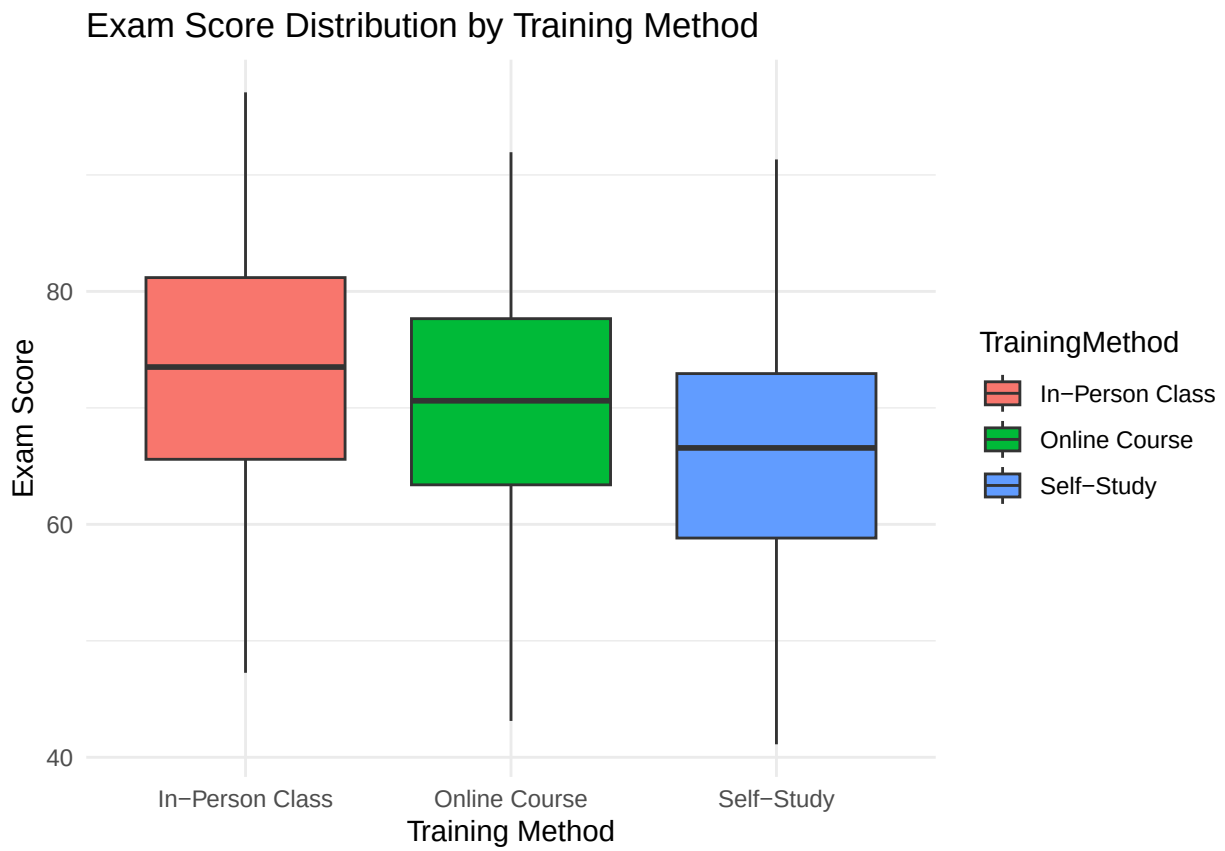
April, 2025

Question 1

(a)

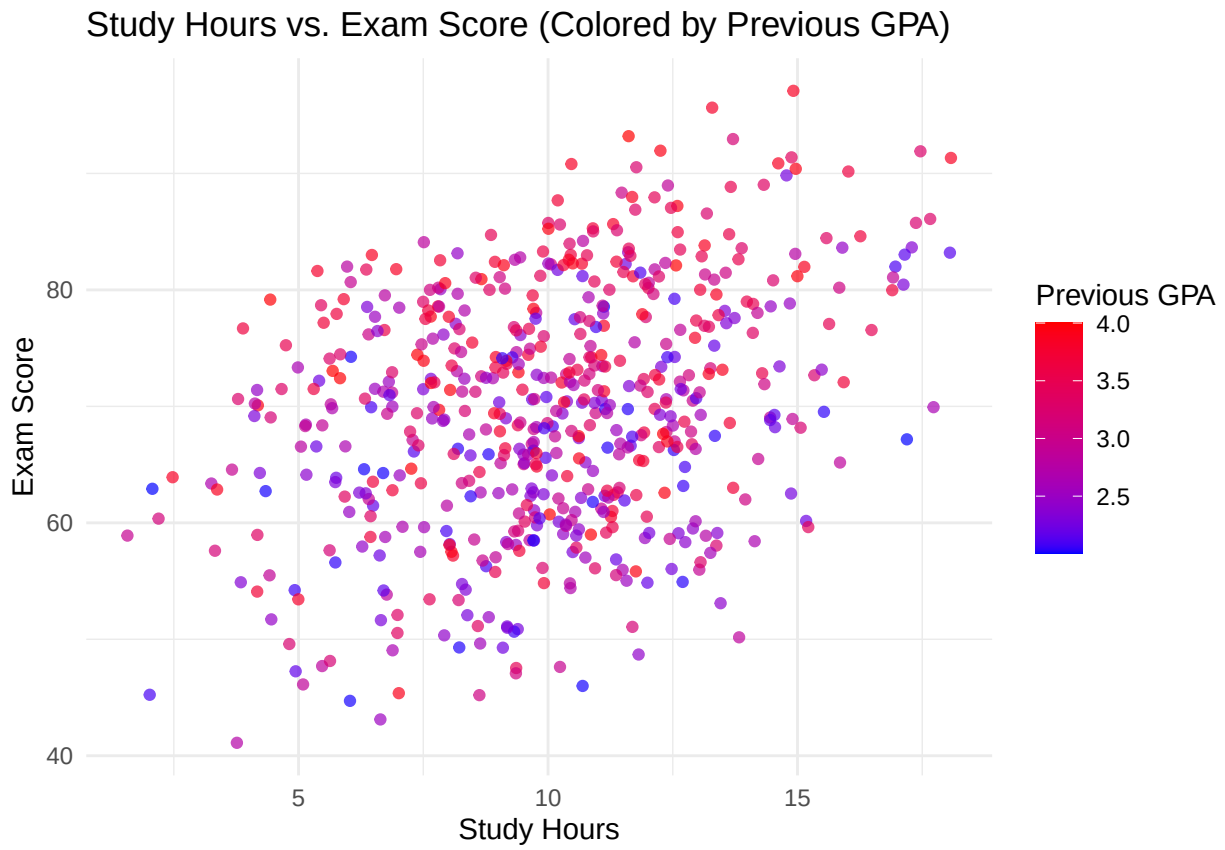
```
student <- read.csv("students_data.csv", header = TRUE)
# head(student)

# Boxplot
ggplot(student, aes(x = TrainingMethod, y = ExamScore, fill = TrainingMethod)) +
  geom_boxplot() +
  labs(title = "Exam Score Distribution by Training Method",
       x = "Training Method",
       y = "Exam Score") +
  theme_minimal()
```



For our boxplot, In-Person Classes hold the highest median score and a relatively wider inter-quartile range. Self-Study Method holdes the lowest median score while Online Courses seem to have a smaller inter-quartile range. Overall, students under the In-person class method seem to perform better on average.

```
# Scatterplot
ggplot(student, aes(x = StudyHours, y = ExamScore, color = PreviousGPA)) +
  geom_point(alpha = 0.7) +
  scale_color_gradient(low = "blue", high = "red") +
  labs(title = "Study Hours vs. Exam Score (Colored by Previous GPA)",
       x = "Study Hours",
       y = "Exam Score",
       color = "Previous GPA") +
  theme_minimal()
```



Overall, we can observe a slight positive correlation between exam scores and study hours. Furthermore, we can see that higher GPA holders (3.5 - 4.0) tend to cluster at the upper right of the graph with high exam score and study hours. On the other hand, lower GPA holders (3.0 - 0) tend to be at the bottom right of the scatterplot whose students have lower exam scores and study hours. Therefore, we can say that students who have higher exam scores.

(b)

```
student$TrainingMethod <- as.factor(student$TrainingMethod)
model <- lm(ExamScore ~ StudyHours + PreviousGPA + TrainingMethod, data = student)
summary(model)
```

```
##
## Call:
## lm(formula = ExamScore ~ StudyHours + PreviousGPA + TrainingMethod,
##     data = student)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1283  -6.3001   0.6668   6.8813  15.6736
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   46.3115     2.2670  20.429  < 2e-16 ***
## StudyHours                     1.0688     0.1199   8.912  < 2e-16 ***
## PreviousGPA                    5.3071     0.6237   8.509  < 2e-16 ***
## TrainingMethodOnline Course   -2.8668     0.8880  -3.228  0.00131 **
## TrainingMethodSelf-Study      -6.6728     0.8913  -7.487 2.56e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.877 on 595 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2639
## F-statistic: 54.68 on 4 and 595 DF,  p-value: < 2.2e-16
```

$ExamScore = 46.3115 + 1.0688 * (StudyHours) + 5.3071 * (PreviousGPA) - 2.8668 * (OnlineCourse) - 6.6728 * (Self - Study) + epsilon$

For Coefficient: Study Hours For an increase of 1 studying hour per week, we expect an increase in exam score by approximately 1.06 points

For Coefficient: PreviousGPA For an increase of 1 Previous GPA point, we expect an increase in exam score by approximately 5.31 points

For Coefficient: TrainingMethod - Online Course For students taking Online Course, they average 2.87 points lower on the exam compared to those who attended In-person classes

For Coefficient: TrainingMethod - Self Study For students who Self Studied, they average 6.67 points lower on the exam compared to those who attended In-person classes

(c) Given: Study Hours per week = 8 Previous GPA = 3.8 Online Course = 1 Self Study = 0

```
46.3115 + 1.0688*(8) + 5.3071*(3.8) - 2.8668*(1) - 6.6728*(0)
```

```
## [1] 72.16208
```

Therefore, this student's expected exam score is ~ 72.16

(d) Null Hypothesis The mean exam scores are equal across all methods $H_0 : \mu_{SelfStudy} = \mu_{Onlineclass} = \mu_{InPerson}$

Alternative Hypothesis At least one training method has a different mean exam score

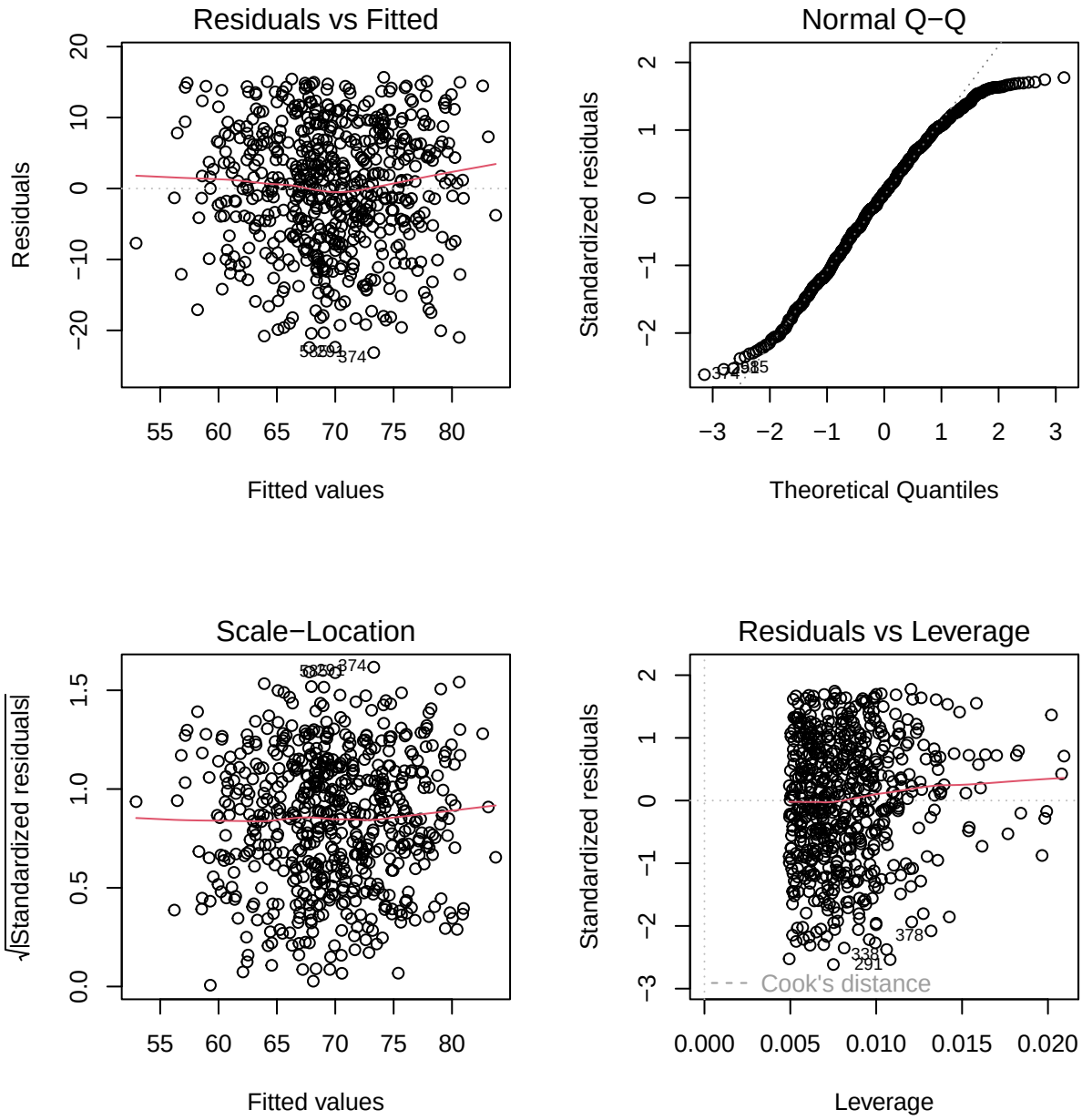
```
student$TrainingMethod <- as.factor(student$TrainingMethod)
anova_model <- aov(ExamScore ~ TrainingMethod, data = student)
summary(anova_model)
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## TrainingMethod   2   4731  2365.5   23.77 1.16e-10 ***
## Residuals      597  59398    99.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we observe a small p-value 1.16e-10 for our categorical variable - Training Method, we can reject the null hypothesis at 0.05 significance level. Therefore, training methods is significant on exam scores.

(e)

```
par(mfrow = c(2, 2))
plot(model)
```



Residual vs Fitted Plot: we can see no patterns which suggests Linearity
and non-funneling and constant variance which suggests Homoscedasticity.

Normal Q-Q plot: We can see that our points roughly follows the 'normal' diagonal line but with heavier tails and skewed, thus suggesting that the residuals may not be normally distributed (CLT Theorem).

(f)

```
int_model <- lm(ExamScore ~ StudyHours * PreviousGPA + TrainingMethod, data = student)
summary(int_model)
```

```
##
```

```
## Call:
## lm(formula = ExamScore ~ StudyHours * PreviousGPA + TrainingMethod,
##     data = student)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.0505  -6.2309   0.5299   7.1292  16.2867
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   51.6493     6.3889   8.084 3.52e-15 ***
## StudyHours                     0.5348     0.6095   0.878  0.38056
## PreviousGPA                    3.4968     2.1196   1.650  0.09952 .
## TrainingMethodOnline Course   -2.8397     0.8887  -3.195  0.00147 **
## TrainingMethodSelf-Study      -6.6750     0.8914  -7.488 2.54e-13 ***
## StudyHours:PreviousGPA         0.1804     0.2018   0.894  0.37187
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.879 on 594 degrees of freedom
## Multiple R-squared:  0.2698, Adjusted R-squared:  0.2636
## F-statistic: 43.89 on 5 and 594 DF,  p-value: < 2.2e-16
```

```
summary(model)
```

```
##
## Call:
## lm(formula = ExamScore ~ StudyHours + PreviousGPA + TrainingMethod,
##     data = student)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1283  -6.3001   0.6668   6.8813  15.6736
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   46.3115     2.2670  20.429  < 2e-16 ***
## StudyHours                     1.0688     0.1199   8.912  < 2e-16 ***
## PreviousGPA                    5.3071     0.6237   8.509  < 2e-16 ***
## TrainingMethodOnline Course   -2.8668     0.8880  -3.228  0.00131 **
## TrainingMethodSelf-Study      -6.6728     0.8913  -7.487 2.56e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.877 on 595 degrees of freedom
## Multiple R-squared:  0.2688, Adjusted R-squared:  0.2639
## F-statistic: 54.68 on 4 and 595 DF,  p-value: < 2.2e-16
```

```
AIC(model, int_model)
```

```
##           df      AIC
## model      6 4329.926
## int_model  7 4331.120
```

Looking at the p-values for the interaction term, we can see that it is a higher pvalue at 0.3718 which is statistically insignificant and that adding the interaction term may not be necessary.

Furthermore, we can see that the Adjusted R-Squared between the two models are very similar with the interaction model being higher by 0.0003.

We can also examine their AIC values to see the better model, and we can observe a lower AIC value for our linear model without interaction, thus suggesting that the interaction model may be a more insignificant explanatory variable.

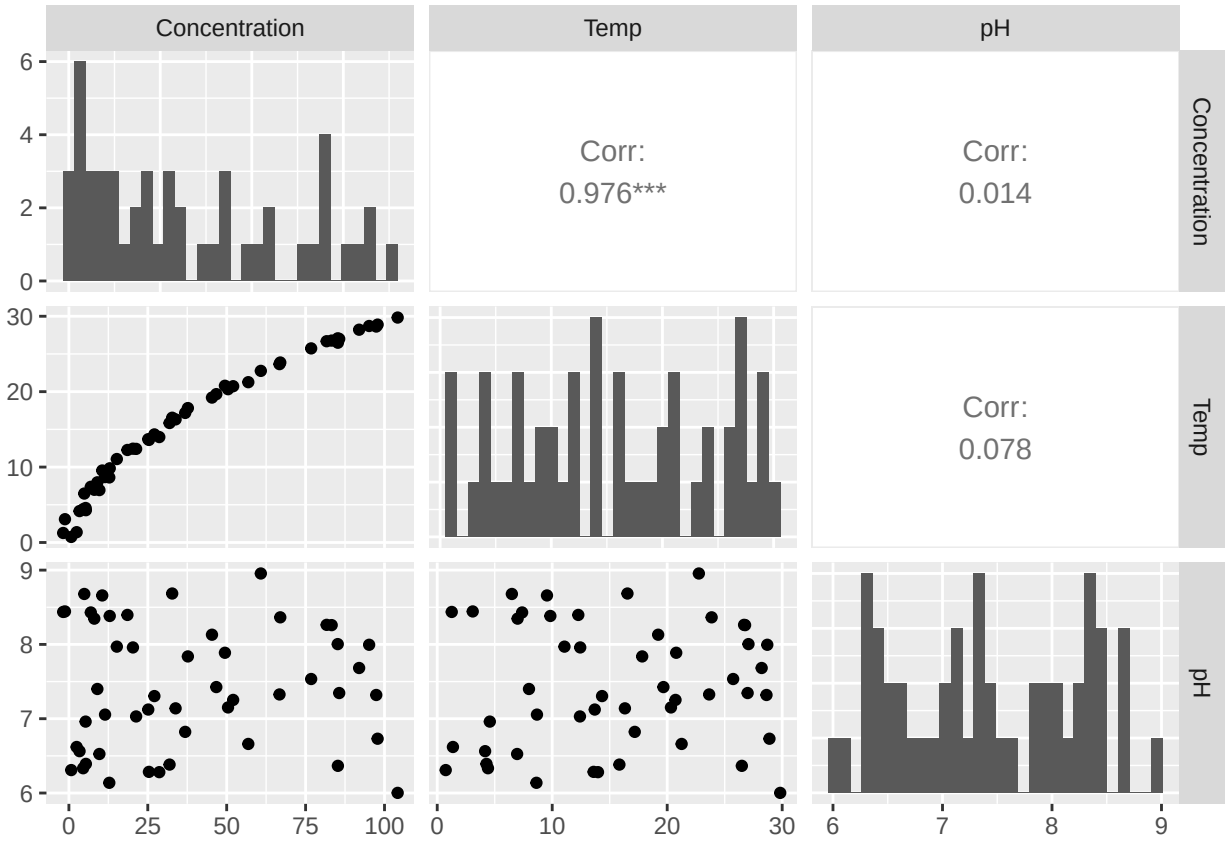
Question 2

(a)

```
data <- read.csv("data.csv")

# Scatterplot Matrix
ggpairs(data,
        diag = list(continuous = "barDiag"),
        lower = list(continuous = "points"),
        upper = list(continuous = "cor"))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We can observe a strong correlation between Temperature and Concentration with 0.976 while the other

relationship between other variables seems weaker with less than 0.1. Looking at the scatter plots, we can observe a linear positive increasing pattern between Temperature and Concentration while the other scatter plots seem more Non-linear without any obvious pattern.

(b)

```
model_b <- lm(Concentration ~ Temp + pH, data = data)

summary(model_b)
```

```
##
## Call:
## lm(formula = Concentration ~ Temp + pH, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3519  -6.1366  -0.9104   5.2172  14.0833
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.3701     9.1422    0.04   0.9679
## Temp          3.6394     0.1145   31.78   <2e-16 ***
## pH           -2.4722     1.2181   -2.03   0.0481 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.055 on 47 degrees of freedom
## Multiple R-squared:  0.9555, Adjusted R-squared:  0.9537
## F-statistic: 505.1 on 2 and 47 DF,  p-value: < 2.2e-16
```

The model equation is: $/ Concentration = 0.3701 + 3.6394(Temp) - 2.4722(pH) $

Coefficient Interpretations: Temperature: For an increase in 1 Degree Celcius, we expect an increase in pollutant concentration by 3.6394 units

pH: For an increase in 1 pH Level, we expect a decrease in pollutant concentration by 2.4722 units

(c) A linear model seems appropriate because the scatter plot between Temperature and Concentration displays a strong positive linear pattern. Our histograms also do not have any obvious outliers that would suggest a more complex model. Also looking into the p-value and Adjusted R-squared of our linear model, I believe that it is adequate in capturing the relationship.

(d) Since we have already fitted an Additive linear regression, we will extend the model by applying an interaction term between Temperature and pH :

```
model_d <- lm(Concentration ~ Temp * pH, data = data)

summary(model_d)
```

```
##
## Call:
## lm(formula = Concentration ~ Temp * pH, data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.285  -6.089  -1.069   5.167  14.445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.93433   16.67138   0.176  0.86106
## Temp         3.45312    1.01500   3.402  0.00139 **
## pH          -2.82304    2.26291  -1.248  0.21852
## Temp:pH      0.02532    0.13704   0.185  0.85422
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.129 on 46 degrees of freedom
## Multiple R-squared:  0.9556, Adjusted R-squared:  0.9527
## F-statistic: 329.8 on 3 and 46 DF,  p-value: < 2.2e-16
```

The model equation is: $/ Concentration = 2.93433 + 3.45312(Temp) - 2.82304(pH) + 0.02532 (Temp)(pH) $

Coefficient Interpretations: Temperature: For an increase in 1 Degree Celcius, we expect an increase in pollutant concentration by 3.45312 units

pH: For an increase in 1 pH Level, we expect a decrease in pollutant concentration by 2.82304 units

Interaction Term: For an increase by 1 unit in both Temperature and pH Level, we expect an increase in pollutant concentration by 0.02532 units

Given the interaction's p-value to be quite high with 0.85422, that suggests that the interaction term is not statistically significant at confidence level 5%. This suggests that including the interaction may not meaningfully improve the model.

(e)

```r
# R squared
summary(model_b)$r.squared
```

```
## [1] 0.9555445
```

```r
summary(model_d)$r.squared
```

```
## [1] 0.9555774
```

```r
# Adjusted R squared
summary(model_b)$adj.r.squared
```

```
## [1] 0.9536527
```

```r
summary(model_d)$adj.r.squared
```

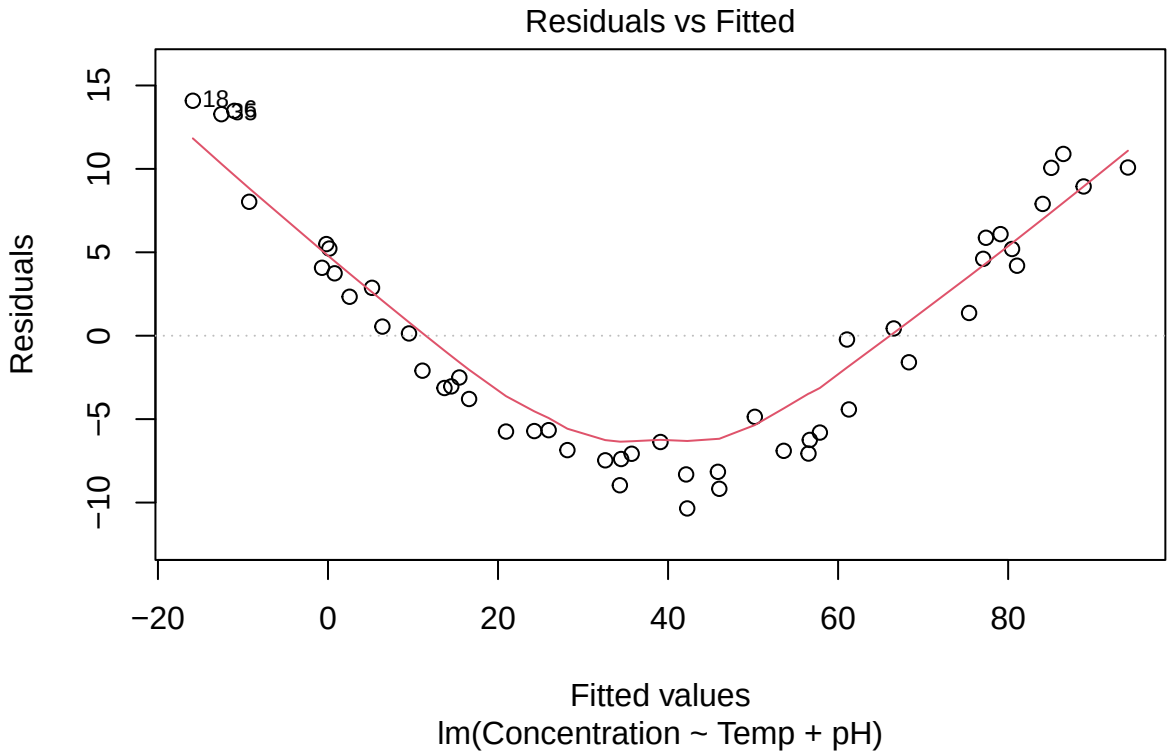
```
## [1] 0.9526803
```

In terms of R squared, both models explained about 95.55% of the variation with the interaction model explaining better than our additive model by slightly more with a difference of 3.29e-05.

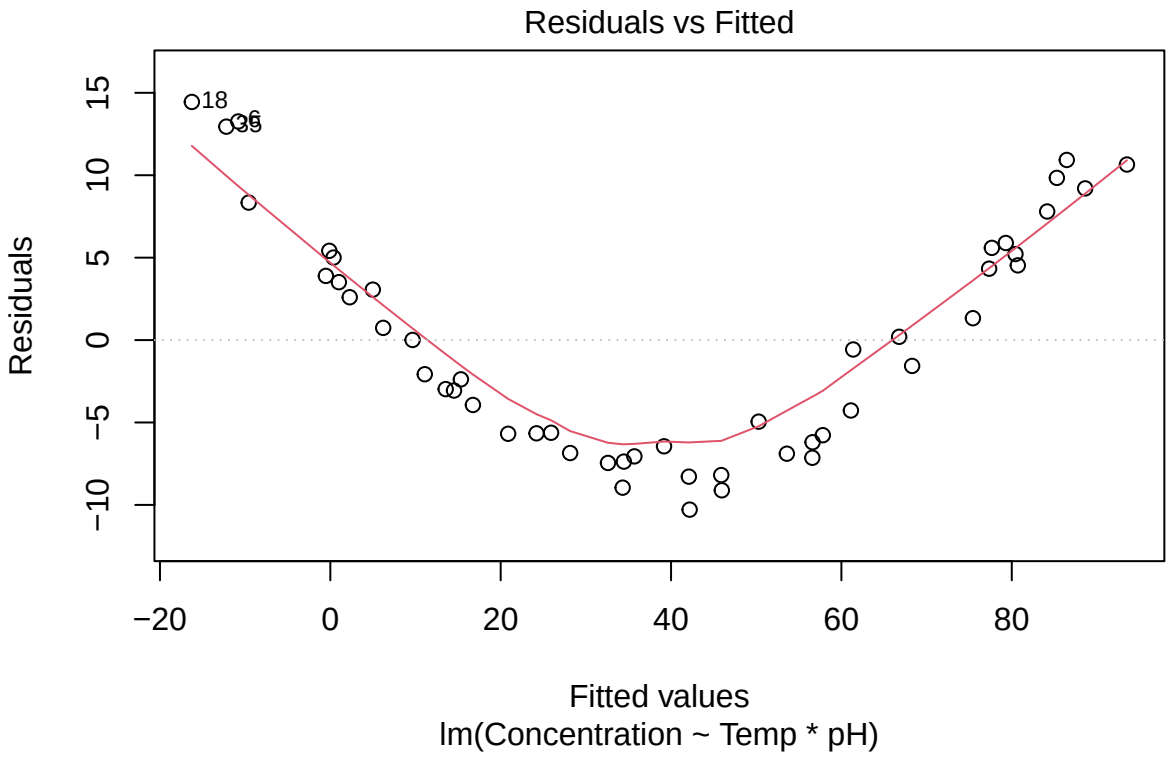
The addition of an interaction term in Model D did not provide any significant improvement in terms of model performance with a slightly lower adjusted R squared value by 0.0009724 compared to the additive model. Also given that the p-value of the interaction term was not statistically significant, Model B (additive) should be the preferred model.

(f) Using the plot() function that includes the residual plot :

```
plot(model_b, which = 1)
```




```
plot(model_d, which = 1)
```



For both additive and multiplicative model, we can observe a violation of homoscedasticity because the variance is not constant throughout and shows a pattern which also violates the linearity-assumption as we can see a visible pattern (curve pattern).

Question 3

(a)

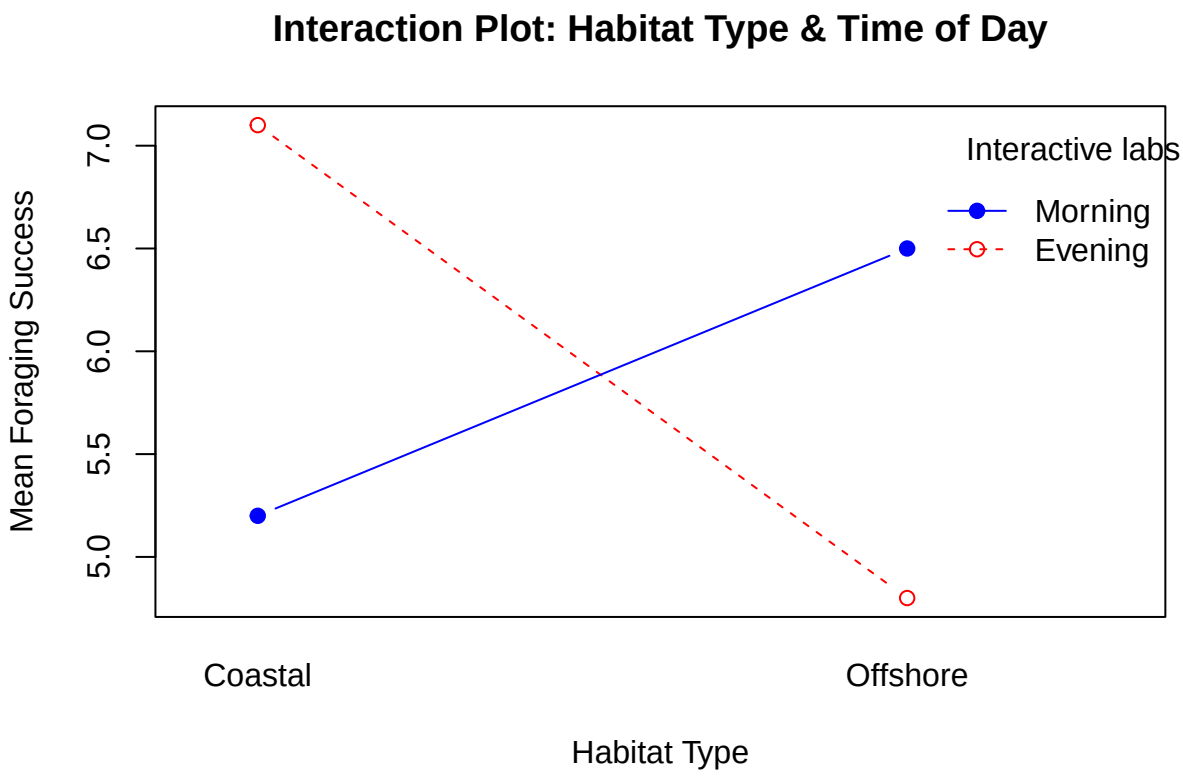
```
# Creating data frame
habitat <- c("Coastal", "Coastal", "Offshore", "Offshore")
time <- c("Morning", "Evening", "Morning", "Evening")
mean_success <- c(5.2, 7.1, 6.5, 4.8)

forage_df <- data.frame(Habitat = habitat, Time = time, MeanSuccess = mean_success)
forage_df
```

```
##    Habitat    Time MeanSuccess
## 1  Coastal Morning         5.2
## 2  Coastal Evening         7.1
## 3 Offshore Morning         6.5
## 4 Offshore Evening         4.8
```

```
# Create interaction plot (From Worksheet 17)
interaction.plot(x.factor = forage_df$Habitat,
                 trace.factor = forage_df$Time,
```

```
                response = forage_df$MeanSuccess,
                type = "b",
                main = "Interaction Plot: Habitat Type & Time of Day",
                xlab = "Habitat Type",
                ylab = "Mean Foraging Success",
                trace.label = "Interactive labs",
                pch=c(1,19), col = c("red", "blue"))
```



The cross lines in our interaction plot suggests that there exist an interaction effect. The plot shows that the average foraging success is higher for morning offshore foraging and evening coastal foraging.

(b) Main Effect (Habitats) $/ H_0 $: There is no difference in mean foraging success between coastal and offshore habitats. $/ H_a $: There is a difference in mean foraging success between coastal and offshore habitats.

Main Effect (Time of Day) $/ H_0 $: There is no difference in mean foraging success between morning and evening. $/ H_a $: There is a difference in mean foraging success between morning and evening.

Interaction Effect $/ H_0 $: There is no interaction between habitat type and time of day $/ H_a $: There is an interaction between habitat type and time of day

(c)

Interaction Term (7.1 - 5.2) - (4.8 - 6.5) = 1.9 + 1.7 = 3.6

(d)

```r
# Sum of Squares
SS_habitat <- 10.3
SS_time <- 8.7
SS_interaction <- 12.5
SS_total <- 46.7
SS_error <- SS_total - SS_habitat - SS_time - SS_interaction

# Degree of Freedom
df_habitat = df_time = df_interaction = 1
df_error = 36
df_total = df_habitat + df_time + df_interaction + df_error

# Mean Square
MS_habitat <- SS_habitat / df_habitat
MS_time <- SS_time / df_time
MS_interaction <- SS_interaction / df_interaction
MS_error <- SS_error / df_error

# F value
F_habitat <- MS_habitat / MS_error
F_time <- MS_time / MS_error
F_interaction <- MS_interaction / MS_error

anova_table <- data.frame(
  Source = c("Habitat", "Time", "Interaction", "Error", "Total"),
  SS = c(SS_habitat, SS_time, SS_interaction, SS_error, SS_total),
  df = c(df_habitat, df_time, df_interaction, df_error, df_total),
  MS = c(MS_habitat, MS_time, MS_interaction, MS_error, NA),
  F = c(F_habitat, F_time, F_interaction, NA, NA)
)
anova_table
```

```
##        Source   SS df         MS        F
## 1     Habitat 10.3  1 10.3000000 24.39474
## 2        Time  8.7  1  8.7000000 20.60526
## 3 Interaction 12.5  1 12.5000000 29.60526
## 4       Error 15.2 36  0.4222222       NA
## 5       Total 46.7 39         NA       NA
```

(e)

```r
qf(0.95, df1 = 1, df2 = 36)
```

```
## [1] 4.113165
```

Since all F-values (Habitat 24, Time 20, Interaction 29) are greater than 4.11, so all effects are statistically significant at the 0.05 level.

Question 4

(a)

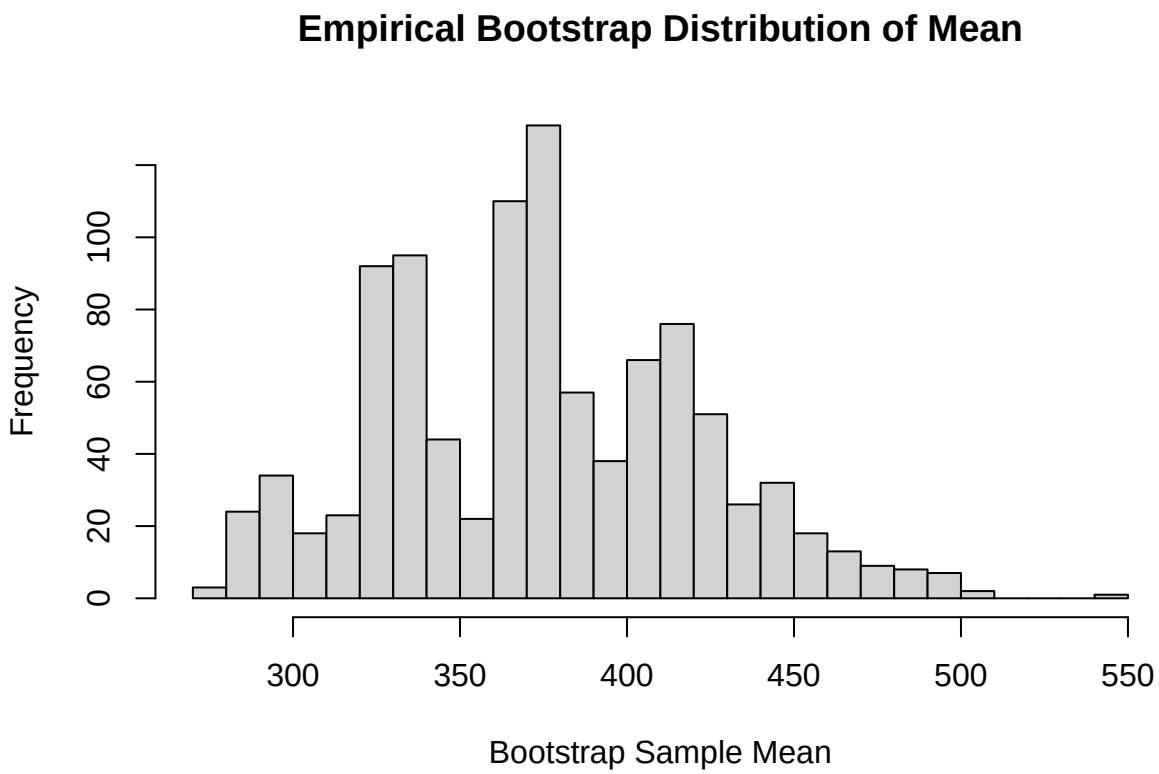
```r
set.seed(123)
coffee <- c(564.4, 498.2, 259.2, 303.3, 299.5, 307.2)

bootstrap_means <- numeric(1000)

# Bootstrapping
for (i in 1:1000) {
  bootstrap_sample <- sample(coffee, size = 6, replace = TRUE)
  bootstrap_means[i] <- mean(bootstrap_sample) # Compute the mean of the bootstrap sample
}
head(bootstrap_means)
```

```
## [1] 354.8667 340.1333 356.6167 414.8333 373.2500 370.6833
```

```r
# Sampling Distribution
hist(bootstrap_means,
     main = "Empirical Bootstrap Distribution of Mean",
     xlab = "Bootstrap Sample Mean",
     breaks = 30)
```



(b)

```r
sd(bootstrap_means)
```

```
## [1] 46.89316
```

The bootstrap estimate's standard deviation tell us the variability of the sample mean across the 1000 bootstrap samples. It is significant to tell if our sample mean is relatively stable (smaller sd) and if it is consistent across different samples from the population.

(c)

```r
sorted_means <- sort(bootstrap_means)

lower_bound <- quantile(sorted_means, 0.025)
lower_bound
```

```
##     2.5%
## 289.2667
```

```r
upper_bound <- quantile(sorted_means, 0.975)
upper_bound
```

```
##    97.5%
## 470.8187
```

There is a 95% chance that the true population mean lies within 289.2667 and 470.8187.

(d) Null Hypothesis: The Mean Caffeine Content = 300mg Alternative Hypothesis: The Mean Caffeine Content != 300 mg

```r
observed_mean <- mean(coffee)

# Using our previous question's bootstrapping
observed_diff <- abs(observed_mean - 300)
extreme_count <- sum( abs(bootstrap_means - 300) >= observed_diff )
p_value <- extreme_count / 1000
p_value
```

```
## [1] 0.5
```

Since our p-value is at 0.5, we reject the null hypothesis at 5% significance level. Hence, there is insufficient evidence to conclude that the mean caffeine content differs from 300 mg.

(e)

```r
t_test_result <- t.test(coffee, mu = 300)
t_test_result
```

```
##
##  One Sample t-test
##
## data:  coffee
## t = 1.395, df = 5, p-value = 0.2218
## alternative hypothesis: true mean is not equal to 300
## 95 percent confidence interval:
##  239.3527 504.5806
## sample estimates:
## mean of x
##  371.9667
```

Using the one-sample t-test, we are assuming that the data follows a normal distribution and comparing the observed sample mean to the null hypothesis value. Although our bootstrapping test does not have much assumptions, our conclusion from our p-value does not differ with the one sample t-test where we reject the null hypothesis.

(f) As the sample size increases, the bootstrap p-value will likely become more precise because our interval of bootstrap sample means will have lower variability.

Given that the Standard Error (SE) = s / sqrt(n), increasing the sample size n will decrease the Standard Error. Hence, our test statistic would be larger as our Smaller SE would produce a more precise estimate. This would also result in a smaller p-value for the t-test which would make us more likely to fail to reject the Null Hypothesis. Furthermore, a larger sample size allows us to apply Central Limit Theorem, where our sampling distribution of the sample mean approaches a normal distribution.