# SamanthaSedar_A07_GLM.Pmd

Samantha Sedar

Fall 2023

## OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

## Directions

1. Rename this file `<FirstLast>_A07_GLMs.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

## Set up your session

1. Set up your session. Check your working directory. Load the tidyverse, agricolae and other needed packages. Import the *raw* NTL-LTER raw data file for chemistry/physics (`NTL-LTER_Lake_ChemistryPhysics_Raw.csv`). Set date columns to date objects.

2. Build a ggplot theme and set it as your default theme.

```
#1 importubg necessary

library(tidyverse); library(lubridate); library(here); library(cowplot); library(agricolae); library(dp
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
## here() starts at /home/guest
##
##
## Attaching package: 'cowplot'
##
```

```
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

```
## [1] "/home/guest"
```

```
NTL_LTER <-
  read.csv(
    "~/EDE_Fall2023/Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
    stringsAsFactors = TRUE)

NTL_LTER$sampledate <- as.Date(NTL_LTER$sampledate, format = "%m/%d/%Y")

#2

mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

## Simple regression

Our first research question is: Does mean lake temperature recorded during July change with depth across all lakes?

3. State the null and alternative hypotheses for this question: > Answer: H0: Mean lake temperature recorded during July does not change with depth across all lakes (alpha and beta are equal to zero) Ha: Mean lake temperature recorded during July changes with depth across all lakes (alpha and beta are not equal to zero)
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4

'Use pipe to filter and select'
```

```
## [1] "Use pipe to filter and select"
```

```
NTL_Processed <-
  NTL_LTER %>%
  filter(format(sampledate, "%m") == "07")%>%
  select(lakename,
```

```
         year4,
        daynum,
        depth,
        temperature_C) %>%
  na.omit()

#5

'plot using lm'
```

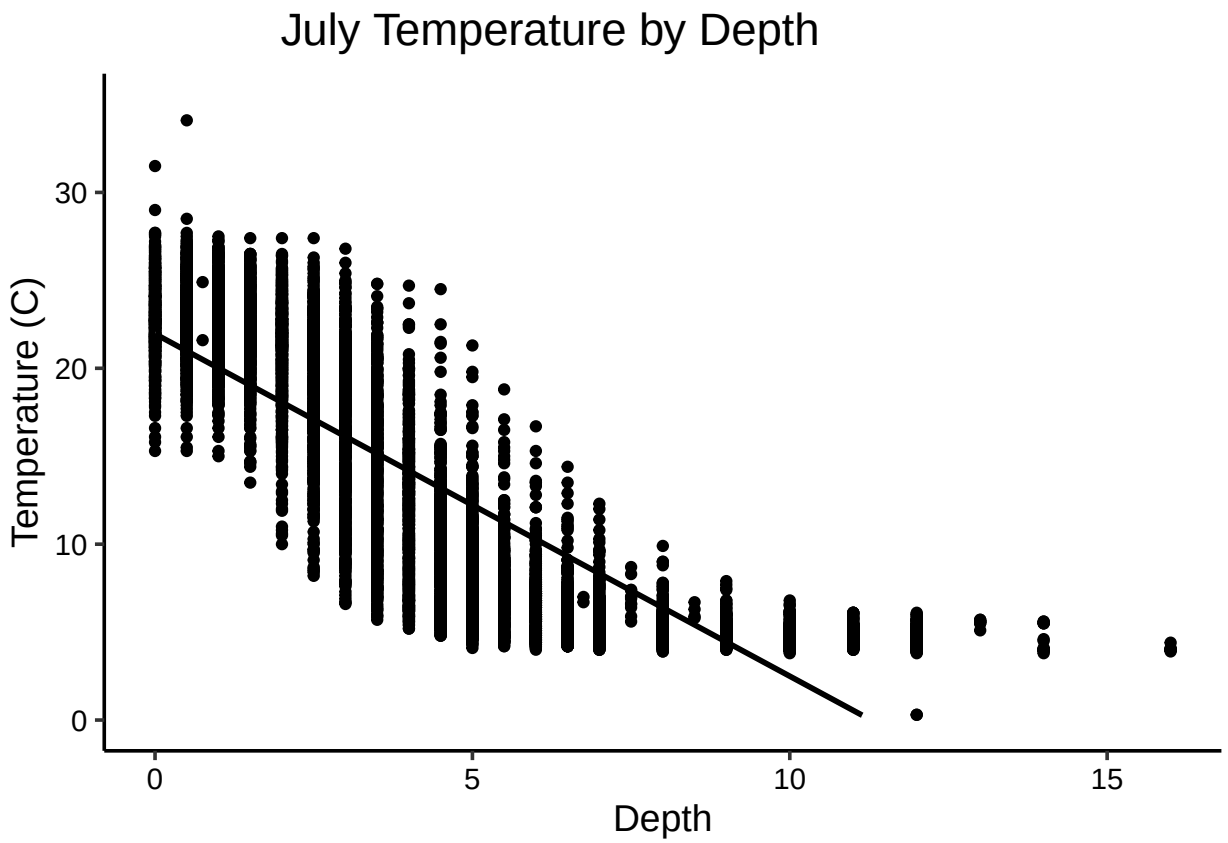
```
## [1] "plot using lm"
```

```
NTL_Processed_plot <- ggplot(NTL_Processed,
                             aes(x=depth,
                             y=temperature_C))+
  ylim(0, 35)+
  geom_point()+
  geom_smooth(method = lm, color="black")+
  labs(title = "              July Temperature by Depth",
       x = "Depth",
       y = "Temperature (C)")

print(NTL_Processed_plot)
```

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

## Warning: Removed 24 rows containing missing values (`geom_smooth()`).
```




6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

   Answer: The figure indicates that there is a negative relationship between temperature and depth. That is, in the month of July, the temperature tends to decrease with an increase in depth, most prominent between 1 and 5 meters.

7. Perform a linear regression to test the relationship and display the results

```
#7

'Regression method one'
```

```
## [1] "Regression method one"
```

```
temperature.regression <-
  lm(NTL_Processed$temperature_C ~
       NTL_Processed$depth)
  summary(temperature.regression)
```

```
##
## Call:
## lm(formula = NTL_Processed$temperature_C ~ NTL_Processed$depth)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         21.95597    0.06792   323.3   <2e-16 ***
## NTL_Processed$depth -1.94621    0.01174  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 on 9726 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7387
## F-statistic: 2.75e+04 on 1 and 9726 DF,  p-value: < 2.2e-16
```

```
'Regression method two'
```

```
## [1] "Regression method two"
```

```
temperature.regression <-
  lm(data = NTL_Processed, temperature_C ~ depth)
  summary(temperature.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_Processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.95597    0.06792   323.3   <2e-16 ***
## depth       -1.94621    0.01174  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 on 9726 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7387
## F-statistic: 2.75e+04 on 1 and 9726 DF,  p-value: < 2.2e-16
```

8. Interpret your model results in words. Include how much of the variability in temperature is explained by changes in depth, the degrees of freedom on which this finding is based, and the statistical significance of the result.

   Answer: Our results illustrate a depth coefficient (slope) of -1.94, signifying a negative relationship between temperature and depth. The p-value is found to be less than 0.05, meaning that the coefficient is statsistically significantly different than zero and it is worthwhile to estimate temperature based on depth. The r squared value shows that depth can account for 74% of variability in temperature. Based on the slope of the overall dataset, we can see that the temperature decreases by 1.94 per 1m.

## Multiple regression

Let's tackle a similar question from a different approach. Here, we want to explore what might the best set of predictors for lake temperature in July across the monitoring period at the North Temperate Lakes LTER.

9. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature.
10. Run a multiple regression on the recommended set of variables.

```
#9

'Run, use step and summarize AIC'
```

```
## [1] "Run, use step and summarize AIC"
```

```
TEMPAIC <- lm(data = NTL_Processed, temperature_C ~ year4+daynum+depth)
step_model <- step(TEMPAIC)
```

```
## Start:  AIC=26065.53
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                141687 26066
## - year4   1       101 141788 26070
## - daynum  1      1237 142924 26148
## - depth   1    404475 546161 39189
```

```
summary(TEMPAIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_Processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   8.630715   -0.994  0.32044
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.817 on 9724 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7411
## F-statistic:  9283 on 3 and 9724 DF,  p-value: < 2.2e-16
```

```
#10

'Re-run based on prior results, since the AIC is lowest for <none> indicating to use all factors'
```

```
## [1] "Re-run based on prior results, since the AIC is lowest for <none> indicating to use all factors"
```

```
TEMPAIC <- lm(data = NTL_Processed, temperature_C ~ year4+daynum+depth)
summary(TEMPAIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_Processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   8.630715   -0.994  0.32044
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.817 on 9724 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7411
## F-statistic:  9283 on 3 and 9724 DF,  p-value: < 2.2e-16
```

11. What is the final set of explanatory variables that the AIC method suggests we use to predict temperature in our multiple regression? How much of the observed variance does this model explain? Is this an improvement over the model using only depth as the explanatory variable?

    Answer: The AIC suggests that all of the explanatory variables are important and should be used to predict the temperature. The AIC for is the lowest, suggesting none should be removed. Based on the r-squared, we can see that this model explains 74.12% of the variance, which is marginally higher than only using depth, which explains 73.87% of variance.

---

## Analysis of Variance

12. Now we want to see whether the different lakes have, on average, different temperatures in the month of July. Run an ANOVA test to complete this analysis. (No need to test assumptions of normality or similar variances.) Create two sets of models: one expressed as an ANOVA models and another expressed as a linear model (as done in our lessons).

```
#12

'Analysis run with ANOVA test'
```

```
## [1] "Analysis run with ANOVA test"
```

```
NTL.anova <- aov(data = NTL_Processed, temperature_C ~ lakename)
summary(NTL.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
'Analysis run as lm'
```

```
## [1] "Analysis run as lm"
```

```
NTL.anova2 <- lm(data = NTL_Processed, temperature_C ~ lakename+year4+daynum+depth)
summary(NTL.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename + year4 + daynum + depth,
##     data = NTL_Processed)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.892 -3.036 -0.218  2.779 15.284
##
## Coefficients:
##                             Estimate Std. Error  t value Pr(>|t|)
## (Intercept)               43.837042   8.246592    5.316 1.09e-07 ***
## lakenameCrampton Lake      4.762283   0.373374   12.755  < 2e-16 ***
## lakenameEast Long Lake    -1.403887   0.333137   -4.214 2.53e-05 ***
## lakenameHummingbird Lake -4.676827   0.452809  -10.328  < 2e-16 ***
## lakenamePaul Lake          1.062414   0.321114    3.309 0.000941 ***
## lakenamePeter Lake         1.500793   0.320895    4.677 2.95e-06 ***
## lakenameTuesday Lake      -1.314121   0.325965   -4.031 5.59e-05 ***
## lakenameWard Lake         -0.418098   0.457643   -0.914 0.360955
## lakenameWest Long Lake    -0.124557   0.332107   -0.375 0.707631
## year4                     -0.015165   0.004119   -3.682 0.000233 ***
## daynum                     0.040749   0.003989   10.217  < 2e-16 ***
## depth                     -1.964288   0.010899 -180.232  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.524 on 9716 degrees of freedom
## Multiple R-squared:  0.7797, Adjusted R-squared:  0.7794
## F-statistic:  3125 on 11 and 9716 DF,  p-value: < 2.2e-16
```

13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer:

Yes, the p-value for lakename is indicated to be less than 0.05, demonstrating a statistically significant evidence against the null hypothesis. This suggests a significant difference in mean temperature among the lakes.

14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.

'New ggplot with 50% transparency'
```

```
## [1] "New ggplot with 50% transparency"
```

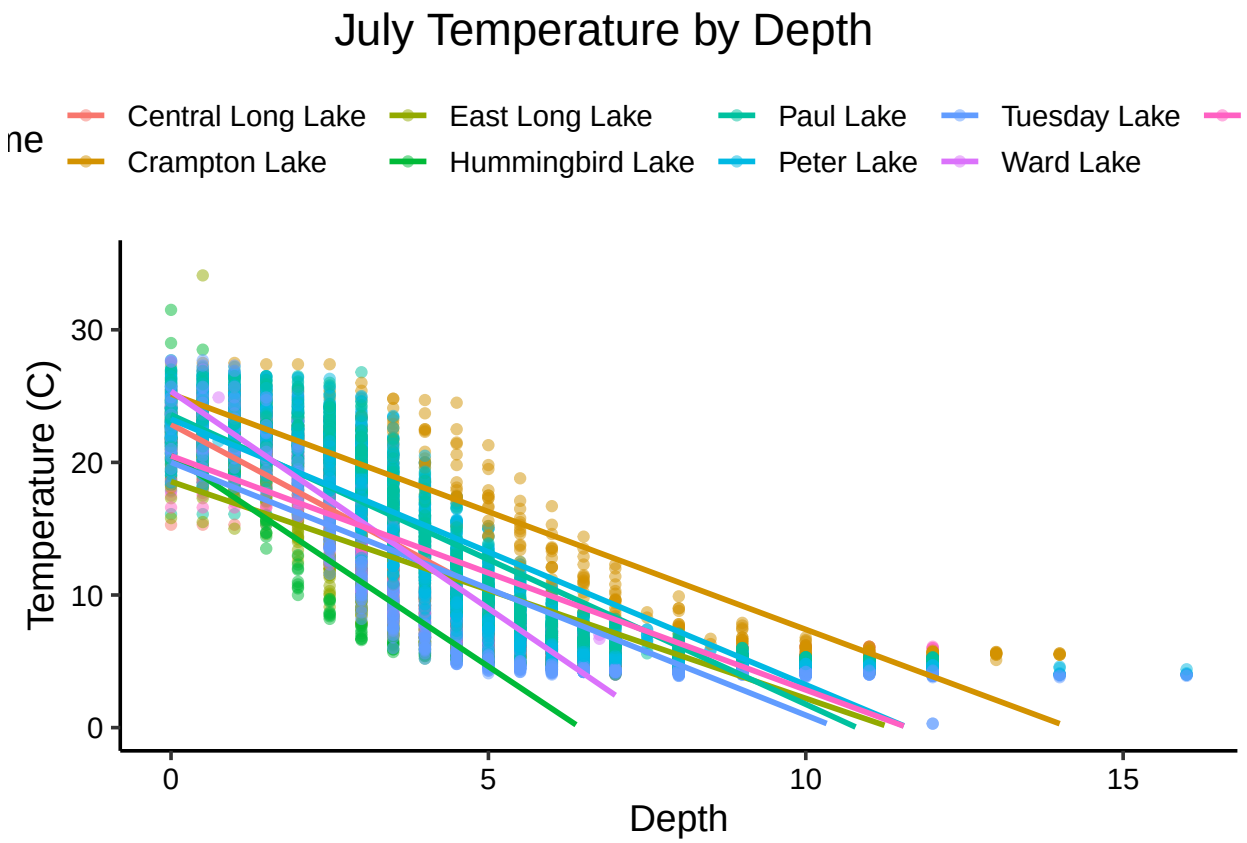
```
NTL_Processed_plot2 <- ggplot(NTL_Processed,
                              aes(x=depth,
                              y=temperature_C, color = lakename))+
  ylim(0, 35)+
  geom_point(alpha = 0.5)+
  geom_smooth(method = lm, se=FALSE, size = 1)+
  labs(title = "                 July Temperature by Depth",
       x = "Depth",
       y = "Temperature (C)")
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
print(NTL_Processed_plot2)
```

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

```
## Warning: Removed 73 rows containing missing values (`geom_smooth()`).
```




15. Use the Tukey's HSD test to determine which lakes have different means.

```
#15 ********
#Tukey's HSD test for lakes
tukey <- HSD.test(NTL.anova, "lakename", group=T)

# Summary of Tukey's HSD results
print(tukey)
```

```
## $statistics
##   MSerror   Df     Mean       CV
##   54.1016 9719 12.72087 57.82135
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387504  0.05
##
## $means
##                   temperature_C      std    r        se Min  Max    Q25   Q50
## Central Long Lake      17.66641 4.196292  128 0.6501298 8.9 26.8 14.400 18.40
## Crampton Lake          15.35189 7.244773  318 0.4124692 5.0 27.5  7.525 16.90
## East Long Lake         10.26767 6.766804  968 0.2364108 4.2 34.1  4.975  6.50
## Hummingbird Lake       10.77328 7.017845  116 0.6829298 4.0 31.5  5.200  7.00
## Paul Lake              13.81426 7.296928 2660 0.1426147 4.7 27.7  6.500 12.40
## Peter Lake             13.31626 7.669758 2872 0.1372501 4.0 27.0  5.600 11.40
```

```
## Tuesday Lake           11.06923 7.698687 1524 0.1884137 0.3 27.7  4.400  6.80
## Ward Lake              14.45862 7.409079  116 0.6829298 5.7 27.6  7.200 12.55
## West Long Lake         11.57865 6.980789 1026 0.2296314 4.0 25.7  5.400  8.00
##                          Q75
## Central Long Lake 21.000
## Crampton Lake     22.300
## East Long Lake    15.925
## Hummingbird Lake  15.625
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    18.800
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31626      c
## West Long Lake         11.57865      d
## Tuesday Lake           11.06923     de
## Hummingbird Lake       10.77328     de
## East Long Lake         10.26767      e
##
## attr(,"class")
## [1] "group"
```

16.From the findings above, which lakes have the same mean temperature, statistically speaking, as Peter Lake? Does any lake have a mean temperature that is statistically distinct from all the other lakes?

Answer:Peter lake is similar to both Ward and Paul Lake, as they are both group c. Given the overlap in groups, none of the lakes is distinct from all other lakes.

17. If we were just looking at Peter Lake and Paul Lake. What's another test we might explore to see whether they have distinct mean temperatures?

Answer: To compare the mean temperatures of Peter Lake and Paul Lake specifically, we can use a two-sample t-test.

18. Wrangle the July data to include only records for Crampton Lake and Ward Lake. Run the two-sample T-test on these data to determine whether their July temperature are same or different. What does the test say? Are the mean temperatures for the lakes equal? Does that match you answer for part 16?

```
# filter the data for Crampton Lake and Ward Lake
crampton_ward <- NTL_Processed %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))
```

```r
# two-sample t-test
t_test <- t.test(data = crampton_ward, temperature_C ~ lakename)
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by lakename
## t = 1.1181, df = 200.37, p-value = 0.2649
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.6821129  2.4686451
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.35189                    14.45862
```

Answer: Since the p-value is greater than 0.05, we fail to reject the null hypothesis, indicating that we do not have sufficient evidence to determine whether the July temperature of the two lakes is significantly different. This is slightly different from the results in part 16. The Tukey's HSD test suggests that Peter Lake's mean temperature is significantly different from all other lakes.