Assignment 7: GLMs (Linear Regressios, ANOVA, & t-tests)

Andreana Chou

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.

```
knitr::opts_chunk$set(warning = FALSE)
#1

#loaded libraries tidyverse, agricolae, ggplot2, lubridate, here
library(tidyverse)
```

```
## -- 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 errors
```

```
library(agricolae)
library(ggplot2)
library(lubridate)
library(here)
```

```
## here() starts at C:/Users/andre/Documents/R Studio Files/EDE_Fall2023
```

```
#imported NTL-LTER chemistry/physics data file
NTL_LTER <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
                     stringsAsFactors = TRUE)

#converted to dates with lubridate
NTL_LTER$sampledate <- mdy(NTL_LTER$sampledate)

#2

#assigned custom theme to my_theme object
my_theme <- theme_classic(base_size = 12) +
  theme(panel.background = element_rect(color = "lightblue", fill = "white"),
        panel.grid.major = element_line(color = "lightblue", linewidth = 0.5),
        legend.title = element_text(color = "black"),
        legend.position = "right")

#set custom theme as default
theme_set(my_theme)
```

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. Ha: Mean lake temperature recorded during July does differ with depth across all lakes.

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.

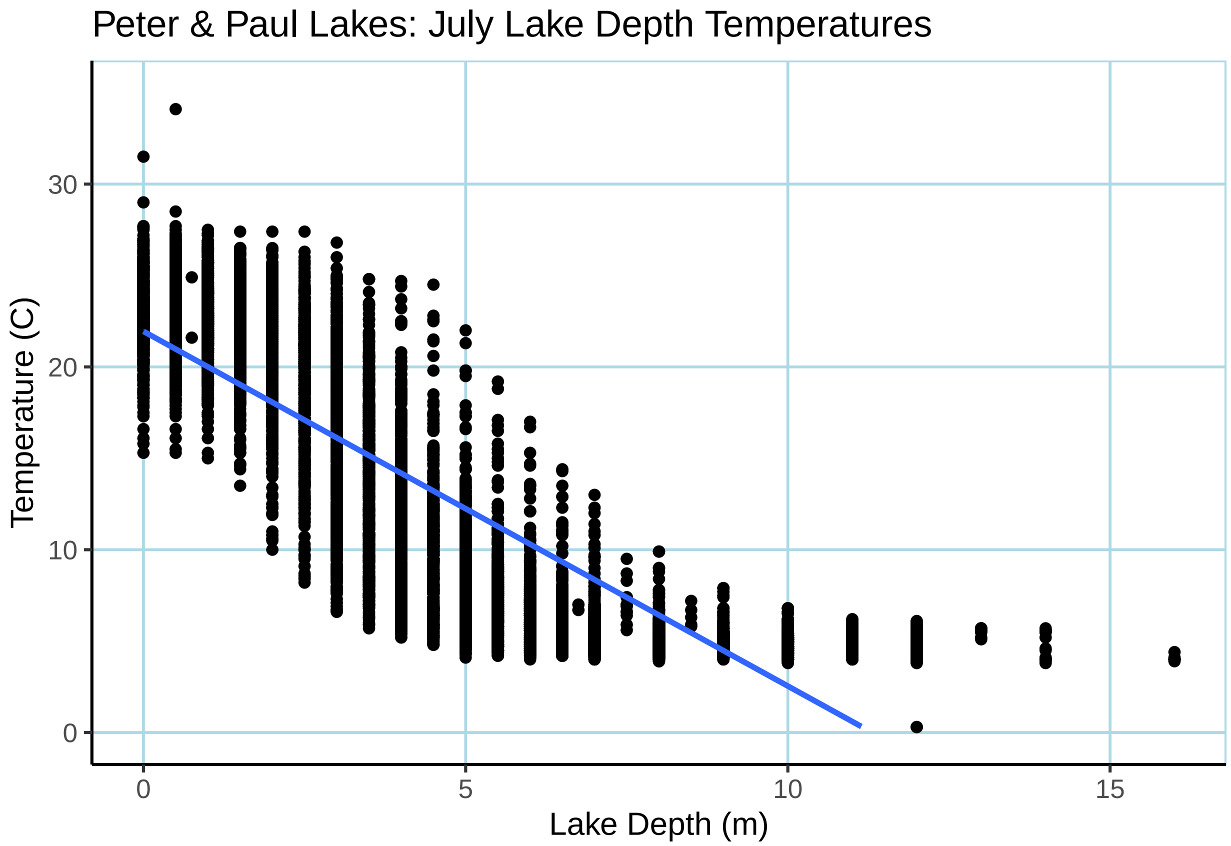
```r
#4

#wrangle data using pipe function, select, filter, and drop_na
NTL_LTER4 <- NTL_LTER %>%
  select("lakename", "year4", "daynum", "depth", "temperature_C") %>%
  filter(between(daynum, 182, 213)) %>%
  drop_na(temperature_C)

#5

#created scatter plot with geom_point, used geom_smooth(method="lm") to create line
#of best fit, ylim to set upper and lower bounds on y-axis
temp_depth_scatter <- NTL_LTER4 %>% ggplot(aes(x=depth, y=temperature_C)) +
  geom_point() +
  geom_smooth(method="lm") + ylim(0, 35) +
  labs(x="Lake Depth (m)", y="Temperature (C)",
       title="Peter & Paul Lakes: July Lake Depth Temperatures")
temp_depth_scatter
```

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



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 suggests temperature will decrease with lake depth, but plateaus after a depth of 10 meters. The distribution of points suggests a linear trend, but mathematical analysis is needed to determine how strong the relationship is between temperature and lake depth.

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

```r
#7

#used lm to perform linear regression on temperature and lake depth variables
temp_depth_regression <- lm(data=NTL_LTER4, depth ~ temperature_C)
summary(temp_depth_regression)
```

```
##
## Call:
## lm(formula = depth ~ temperature_C, data = NTL_LTER4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0778 -1.1159 -0.2393  0.9663  8.0868
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.584075   0.033561   285.6   <2e-16 ***
## temperature_C -0.379735   0.002272  -167.1   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.703 on 9974 degrees of freedom
## Multiple R-squared:  0.7368, Adjusted R-squared:  0.7368
## F-statistic: 2.793e+04 on 1 and 9974 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. Also mention how much temperature is predicted to change for every 1m change in depth.

   Answer: 0.7368 (or 73.68%) of the variability in temperature is explained by changes in depth, based on 9974 degrees of freedom. For every meter increase of depth, the temperature decreases by 0.37 degrees C. These findings are statistically significant because the p-value is less than 0.05.

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

#assigned multi-variable linear regression model to temp_var AIC object
temp_var <- lm(data = NTL_LTER4, temperature_C ~ year4 + daynum + depth)

#performed AIC using AIC object
step(temp_var)
```

```
## Start:  AIC=26781.56
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                146054 26782
## - year4   1       154 146209 26790
## - daynum  1      1582 147636 26887
## - depth   1    414049 560103 40189
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER4)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -14.33180      0.01386      0.04337     -1.94112
```

```
#10

#performed linear regression on AIC's recommended variables
temp_regression <- lm(data = NTL_LTER4, temperature_C ~ year4 + daynum + depth)
summary(temp_regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER4)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.669 -3.014  0.091  2.977 13.606
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -14.331802   8.582522   -1.670  0.09497 .
## year4         0.013861   0.004274    3.243  0.00119 **
## daynum        0.043368   0.004173   10.393  < 2e-16 ***
## depth        -1.941121   0.011545 -168.135  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
```

```
## Residual standard error: 3.827 on 9972 degrees of freedom
## Multiple R-squared:  0.7399, Adjusted R-squared:  0.7398
## F-statistic:  9457 on 3 and 9972 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 final set of explanatory variables suggested by AIC is year4, daynum, and depth. This model explains 0.7398 (73.98%) of the observed variance in temperature. This is a slight (0.3) improvement over the model that uses only depth as the explanatory variable.

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).

```r
#12

#wrangle data by parsing lakename and depth categories, used summarise function to
#create average temperature column
lakes_average <- NTL_LTER4 %>% group_by(lakename, depth) %>%
  summarise(temp_C_average = mean(temperature_C))
```

```
## ‘summarise()‘ has grouped output by ’lakename’. You can override using the
## ‘.groups‘ argument.
```

```r
#performed ANOVA model between average temperature and lake name
lakes_anova <- aov(data=lakes_average, temp_C_average ~ lakename)
summary(lakes_anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lakename      8    727   90.85   1.769 0.0864 .
## Residuals   165   8472   51.34
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```r
#performed linear regression model between temperature and lake name
lakes_anova2 <- lm(data=lakes_average, temp_C_average ~ lakename)
summary(lakes_anova2)
```

```
##
## Call:
## lm(formula = temp_C_average ~ lakename, data = lakes_average)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -9.086 -5.715 -3.066  7.400 13.790
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                16.932      2.266   7.473 4.36e-12 ***
## lakenameCrampton Lake      -2.370      2.697  -0.879  0.38085    
## lakenameEast Long Lake     -6.711      2.775  -2.418  0.01669 *  
## lakenameHummingbird Lake   -6.952      2.925  -2.377  0.01862 *  
## lakenamePaul Lake          -4.409      2.753  -1.602  0.11117    
## lakenamePeter Lake         -4.849      2.681  -1.809  0.07231 .  
## lakenameTuesday Lake       -7.190      2.733  -2.631  0.00932 ** 
## lakenameWard Lake          -2.689      2.856  -0.942  0.34779    
## lakenameWest Long Lake     -5.451      2.775  -1.964  0.05119 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 165 degrees of freedom
## Multiple R-squared:  0.07901,    Adjusted R-squared:  0.03436
## F-statistic: 1.769 on 8 and 165 DF,  p-value: 0.08644
```

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

    Answer: There is not a significant difference in mean temperatures among the lakes because the p-value from both ANOVA and linear models is 0.0864, which is greater than 0.05. Therefore, the null hypothesis that there is no significant difference in mean temperatures, cannot be rejected.

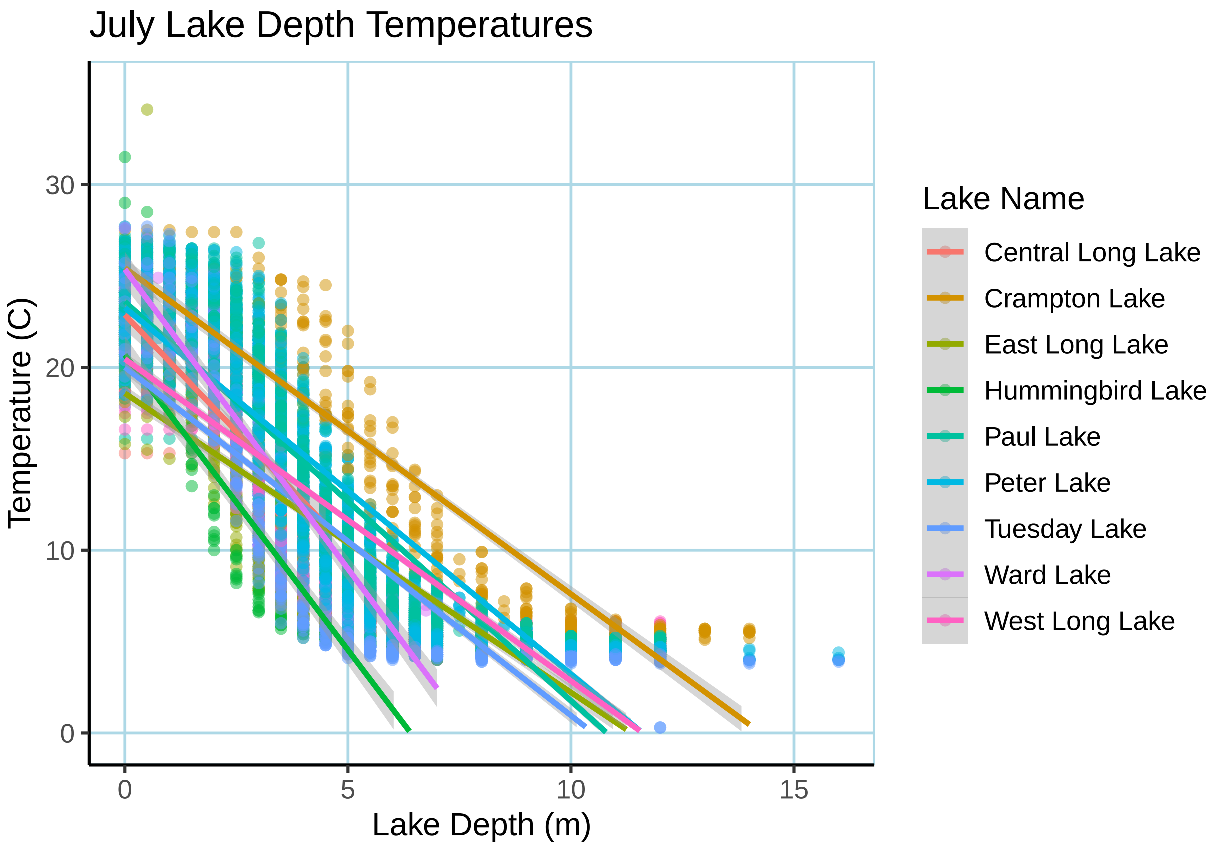
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.

#created scatterplot with geom_point, alpha=0.5 denotes transparency of points,
#geom_smooth(method="lm") to create line of best fit, ylim to set upper
#and lower bounds on y-axis

temp_depth_scatter2 <- NTL_LTER4 %>%
  ggplot(aes(x=depth, y=temperature_C, color=lakename)) +
  geom_point(alpha=0.5) +
  geom_smooth(method="lm") + ylim(0, 35) +
  labs(x="Lake Depth (m)", y="Temperature (C)",
       title="July Lake Depth Temperatures", color="Lake Name")
temp_depth_scatter2
```

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



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

```
#15
#Tukey HSD test contains ANOVA model as argument
TukeyHSD(lakes_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temp_C_average ~ lakename, data = lakes_average)
##
## $lakename
##                                         diff        lwr       upr     p adj
## Crampton Lake-Central Long Lake    -2.3697903 -10.847548  6.107968 0.9937996
## East Long Lake-Central Long Lake   -6.7105173 -15.434058  2.013024 0.2812851
## Hummingbird Lake-Central Long Lake -6.9520503 -16.147470  2.243369 0.3039226
## Paul Lake-Central Long Lake        -4.4091746 -13.063204  4.244855 0.8027693
## Peter Lake-Central Long Lake       -4.8493609 -13.277102  3.578380 0.6763813
## Tuesday Lake-Central Long Lake     -7.1899806 -15.780330  1.400369 0.1819991
## Ward Lake-Central Long Lake        -2.6887605 -11.665210  6.287689 0.9901646
## West Long Lake-Central Long Lake   -5.4509108 -14.174452  3.272630 0.5704391
## East Long Lake-Crampton Lake       -4.3407270 -11.160233  2.478779 0.5450375
## Hummingbird Lake-Crampton Lake     -4.5822599 -11.995844  2.831324 0.5851491
## Paul Lake-Crampton Lake            -2.0393842  -8.769742  4.690974 0.9893893
## Peter Lake-Crampton Lake           -2.4795706  -8.916364  3.957223 0.9531314
## Tuesday Lake-Crampton Lake         -4.8201902 -11.468468  1.828087 0.3608972
## Ward Lake-Crampton Lake            -0.3189702  -7.459148  6.821208 1.0000000
## West Long Lake-Crampton Lake       -3.0811204  -9.900626  3.738385 0.8885949
## Hummingbird Lake-East Long Lake    -0.2415330  -7.934973  7.451907 1.0000000
## Paul Lake-East Long Lake            2.3013427  -4.736093  9.338779 0.9826502
## Peter Lake-East Long Lake           1.8611564  -4.896070  8.618382 0.9943901
## Tuesday Lake-East Long Lake        -0.4794633  -7.438442  6.479515 0.9999998
## Ward Lake-East Long Lake            4.0217568  -3.408581 11.452095 0.7448971
## West Long Lake-East Long Lake       1.2596066  -5.863135  8.382348 0.9997675
## Paul Lake-Hummingbird Lake          2.5428757  -5.071655 10.157407 0.9801945
## Peter Lake-Hummingbird Lake         2.1026893  -5.253646  9.459025 0.9927952
## Tuesday Lake-Hummingbird Lake      -0.2379303  -7.780009  7.304149 1.0000000
## Ward Lake-Hummingbird Lake          4.2632898  -3.715777 12.242356 0.7581114
## West Long Lake-Hummingbird Lake     1.5011395  -6.192301  9.194580 0.9995172
## Peter Lake-Paul Lake               -0.4401864  -7.107432  6.227059 0.9999999
## Tuesday Lake-Paul Lake             -2.7808060  -9.652446  4.090834 0.9380214
## Ward Lake-Paul Lake                 1.7204141  -5.628190  9.069019 0.9981903
## West Long Lake-Paul Lake           -1.0417362  -8.079172  5.995700 0.9999395
## Tuesday Lake-Peter Lake            -2.3406196  -8.924998  4.243759 0.9708745
## Ward Lake-Peter Lake                2.1606004  -4.920119  9.241319 0.9888923
## West Long Lake-Peter Lake          -0.6015498  -7.358776  6.155676 0.9999988
## Ward Lake-Tuesday Lake              4.5012201  -2.772284 11.774724 0.5834925
## West Long Lake-Tuesday Lake         1.7390698  -5.219909  8.698048 0.9971319
## West Long Lake-Ward Lake           -2.7621502 -10.192489  4.668188 0.9619311
```

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: Central Long Lake, Crampton Lake, East Long Lake, Hummingbird Lake, Paul Lake, Ward Lake, West Long Lake all have the statistically same mean temperatures as Peter Lake. There are no lakes with statistically distinct mean temperatures from each other.

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: 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?

```r
#wrangle data by filtering out Crampton Lake and Ward Lake names
crampton_ward <- NTL_LTER4 %>% filter(lakename == c("Crampton Lake", "Ward Lake"))

#t-test function to analyze temperature by lake name
t.test(crampton_ward$temperature_C ~ crampton_ward$lakename, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  crampton_ward$temperature_C by crampton_ward$lakename
## t = 1.15, df = 236, p-value = 0.2513
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.9178817  3.4920576
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.54066                    14.25357
```

Answer: The mean of Crampton Lake (15.54 degrees C) and Ward Lake (14.25 degrees C) are different. However, the t-value is 1.15 with 236 degrees of freedom and a p-value of 0.2513. With a p-value > 0.05, the null hypothesis that "there is no significant difference in Crampton Lake and Ward Lake July temperatures" remains. This aligns with the answer from number 16.