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

Wynona Curaming

Spring 2024

OVERVIEW

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

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
getwd()
```

```
## [1] "/home/guest/ENV 872/EDA_Spring2024"
```

```
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 error
```

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

```
## here() starts at /home/guest/ENV 872/EDA_Spring2024
```

```
library(lubridate)
library(cowplot)
```

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

```
library(ggthemes)
```

```
##
## Attaching package: 'ggthemes'
##
## The following object is masked from 'package:cowplot':
##
##     theme_map
```

```
Lake.data<-read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
                    stringsAsFactors = TRUE)
colnames(Lake.data)
```

```
##  [1] "lakeid"          "lakename"        "year4"           "daynum"
##  [5] "sampledate"      "depth"           "temperature_C"   "dissolvedOxygen"
##  [9] "irradianceWater" "irradianceDeck"  "comments"
```

```
Lake.data$sampledate<-ymd(Lake.data$sampledate)
```

```
## Warning: 33138 failed to parse.
```

```
class(Lake.data$sampledate)
```

```
## [1] "Date"
```

```
#2
my_theme <- theme_base() + theme(line=element_line(color='black',linewidth=1),
                                 plot.background = element_rect(color='beige', fill='beige'),
                                 plot.title = element_text(color='black', size=12),
                                 legend.background = element_rect(
      color='darkgrey',
      fill= 'darkgrey'),legend.text = element_text(color='white',size=12),
      axis.title = element_text(size=12),
      panel.grid.major = element_line(color = "lightgray", linewidth = 0.5),
    panel.grid.minor = element_line(color = "lightgray", linewidth = 0.25))
```

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 in July does not change with depth across all lakes. Ha: Mean lake temperature in July significantly changes with depth across all lakes.
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

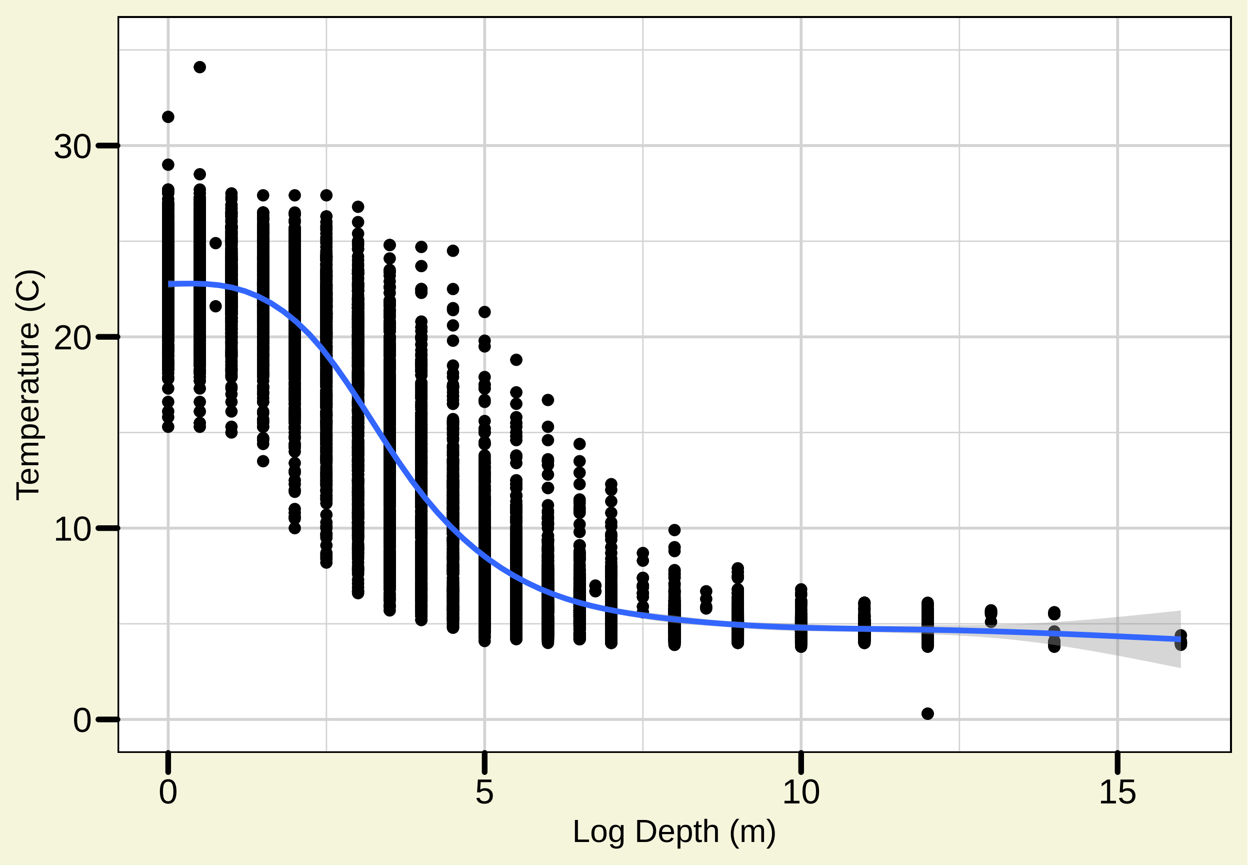
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.

```
#4
Lake.July.df<-Lake.data%>%
  filter(daynum>=182 & daynum<=212)%>%
  select("lakename", "year4", "daynum", "depth", "temperature_C")%>%
  drop_na()

#5
ggplot(Lake.July.df, aes(depth, temperature_C))+
  geom_point()+geom_smooth()+ylim(0,35)+
  labs(x="Log Depth (m)", y="Temperature (C)")+my_theme
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula = ’y ~ s(x, bs = "cs")’
```



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 overall trend is more like an inverse S-shape. As depth increases, temperature decreases but starting around 8 or 9 m depth and beyond, the temperature stays almost constant at around 5 C. The distribution of points suggest that there is more variation in temperature closer to the surface (at lower depths), which makes sense since it would likely be influenced heavily by sea-surface temperatures and weather.

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

```
#7
lm_temp_depth<-lm(temperature_C ~ depth, Lake.data)
summary(lm_temp_depth)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Lake.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7864  -3.1363  -0.1219   3.1815  19.2568
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.986395   0.037166   537.8   <2e-16 ***
## depth       -1.707162   0.006366  -268.2   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.961 on 34754 degrees of freedom
##   (3858 observations deleted due to missingness)
## Multiple R-squared:  0.6742, Adjusted R-squared:  0.6742
## F-statistic: 7.192e+04 on 1 and 34754 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: 67.42% of variability of temperature is explained by changes in depth. The result is statistically significant as the p-value is very small (<0.001). This finding is based on 34754 degrees of freedom. For every 1 meter increase in depth, temperature is expected to drop by -1.7 degrees C. At the surface, temperature is expected to be almost 20 degrees C.

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
colnames(Lake.July.df)
```

```
## [1] "lakename"      "year4"         "daynum"        "depth"
## [5] "temperature_C"
```

```
summary(Lake.July.df)
```

```
##             lakename        year4          daynum          depth
##  Peter Lake    :2892   Min.   :1984   Min.   :182.0   Min.   : 0.000
##  Paul Lake     :2643   1st Qu.:1992   1st Qu.:190.0   1st Qu.: 2.000
##  Tuesday Lake  :1507   Median :1998   Median :197.0   Median : 4.500
##  West Long Lake:1043   Mean   :1999   Mean   :197.3   Mean   : 4.747
##  East Long Lake: 968   3rd Qu.:2006   3rd Qu.:205.0   3rd Qu.: 7.000
##  Crampton Lake : 318   Max.   :2016   Max.   :212.0   Max.   :16.000
##  (Other)       : 351
##  temperature_C
##  Min.   : 0.30
##  1st Qu.: 5.50
##  Median :10.10
##  Mean   :12.71
##  3rd Qu.:20.80
##  Max.   :34.10
##
```

```
lm_temp_depth_AIC<-lm(data=Lake.July.df, temperature_C ~ depth + year4 + daynum)
step(lm_temp_depth_AIC)
```

```
## Start:  AIC=26016.31
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq    RSS   AIC
## <none>                141118 26016
## - year4   1        80 141198 26020
## - daynum  1      1333 142450 26106
## - depth   1    403925 545042 39151
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = Lake.July.df)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##    -6.45556     -1.94726      0.01013      0.04134
```

```
#Answer: AIC is lowest when none is removed so let's include all three explanatory variables to predict

#10
lm_temp_depth_day<-lm(data=Lake.July.df, temperature_C ~ depth + daynum + year4)
summary(lm_temp_depth_day)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum + year4, data = Lake.July.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6517 -2.9937  0.0855  2.9692 13.6171
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -6.455560   8.638808   -0.747   0.4549
## depth       -1.947264   0.011676 -166.782   <2e-16 ***
## daynum       0.041336   0.004315    9.580   <2e-16 ***
## year4        0.010131   0.004303    2.354   0.0186 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.811 on 9718 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7417
## F-statistic:  9303 on 3 and 9718 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 variables are depth, daynum and year4. This model explains 74.17% of observed variance in temperature. This is a 6.75% improvement compared to the model using only depth (67.42).

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
#linear
lm_temp_bylake<-lm(data=Lake.July.df, temperature_C~lakename)
summary(lm_temp_bylake)
```

```
## 
## Call:
## lm(formula = temperature_C ~ lakename, data = Lake.July.df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -10.766  -6.592  -2.692   7.634  23.832 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               17.6731     0.6741  26.218  < 2e-16 ***
## lakenameCrampton Lake     -2.3212     0.7902  -2.938  0.00332 ** 
## lakenameEast Long Lake    -7.4054     0.7143 -10.367  < 2e-16 ***
## lakenameHummingbird Lake  -6.8998     0.9594  -7.192 6.88e-13 ***
## lakenamePaul Lake         -3.8813     0.6891  -5.633 1.82e-08 ***
## lakenamePeter Lake        -4.3710     0.6878  -6.355 2.18e-10 ***
## lakenameTuesday Lake      -6.6073     0.7002  -9.437  < 2e-16 ***
## lakenameWard Lake         -3.2145     0.9594  -3.350  0.00081 ***
## lakenameWest Long Lake    -6.0876     0.7115  -8.556  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 7.353 on 9713 degrees of freedom
## Multiple R-squared:  0.03883,    Adjusted R-squared:  0.03803 
## F-statistic: 49.04 on 8 and 9713 DF,  p-value: < 2.2e-16
```

```
#anova
anova_temp_lake<-aov(data=Lake.July.df, temperature_C~lakename)
summary(anova_temp_lake)
```

```
##                 Df Sum Sq Mean Sq F value Pr(>F)    
## lakename         8  21214  2651.8   49.04 <2e-16 ***
## Residuals     9713 525188    54.1                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

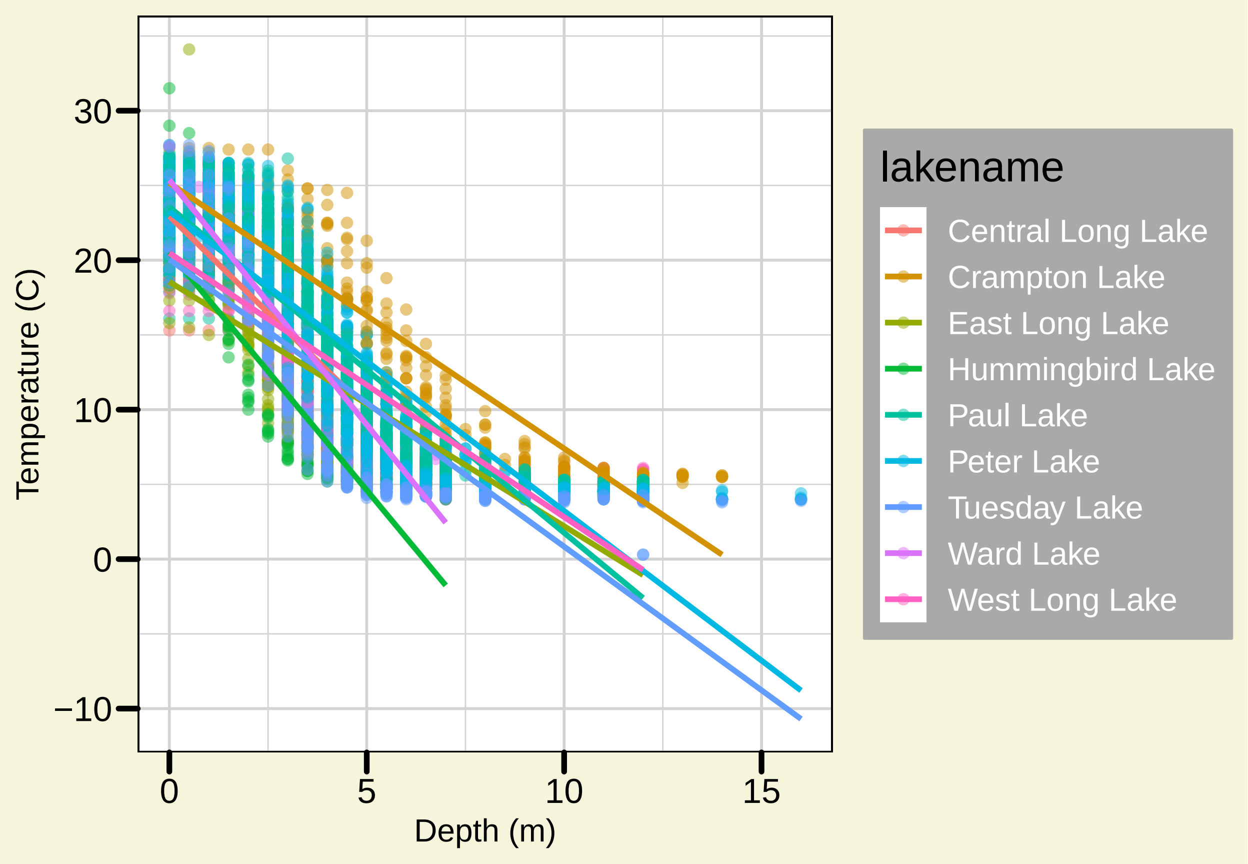
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: We reject the null hypothesis since the p-value is <2e-16. There is a significant difference in mean temperature among the lakes. We also find a low R-squared value which suggests that only 3.89% of July lake temperature variability can be explained by the lake itself. The F-statistic is 49.04.

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.
ggplot(Lake.July.df, aes(x=depth, y=temperature_C, color=lakename))+
geom_point(alpha=0.5)+
geom_smooth(method="lm", se= FALSE)+
labs(x="Depth (m)", y= "Temperature (C)")+my_theme
```

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



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

```
#15
library(multcompView)
TukeyHSD(anova_temp_lake)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = Lake.July.df)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.3212225 -4.7727515  0.1303066 0.0801309
## East Long Lake-Central Long Lake   -7.4054440 -9.6215318 -5.1893561 0.0000000
## Hummingbird Lake-Central Long Lake -6.8998334 -9.8763946 -3.9232722 0.0000000
## Paul Lake-Central Long Lake        -3.8813120 -6.0191419 -1.7434822 0.0000007
## Peter Lake-Central Long Lake       -4.3710346 -6.5048955 -2.2371736 0.0000000
## Tuesday Lake-Central Long Lake     -6.6072831 -8.7795517 -4.4350145 0.0000000
## Ward Lake-Central Long Lake        -3.2144886 -6.1910498 -0.2379273 0.0229685
## West Long Lake-Central Long Lake   -6.0875867 -8.2949346 -3.8802388 0.0000000
## East Long Lake-Crampton Lake       -5.0842215 -6.5587481 -3.6096949 0.0000000
## Hummingbird Lake-Crampton Lake     -4.5786109 -7.0531008 -2.1041211 0.0000003
## Paul Lake-Crampton Lake            -1.5600896 -2.9141574 -0.2060217 0.0106305
## Peter Lake-Crampton Lake           -2.0498121 -3.3976050 -0.7020192 0.0000841
## Tuesday Lake-Crampton Lake         -4.2860606 -5.6938725 -2.8782488 0.0000000
## Ward Lake-Crampton Lake            -0.8932661 -3.3677559  1.5812237 0.9713958
## West Long Lake-Crampton Lake       -3.7663643 -5.2277226 -2.3050060 0.0000000
## Hummingbird Lake-East Long Lake     0.5056106 -1.7358512  2.7470723 0.9988025
## Paul Lake-East Long Lake            3.5241319  2.6670727  4.3811912 0.0000000
## Peter Lake-East Long Lake           3.0344094  2.1872987  3.8815201 0.0000000
## Tuesday Lake-East Long Lake         0.7981609 -0.1415120  1.7378337 0.1721160
## Ward Lake-East Long Lake            4.1909554  1.9494937  6.4324171 0.0000002
## West Long Lake-East Long Lake       1.3178572  0.2997124  2.3360021 0.0019544
## Paul Lake-Hummingbird Lake          3.0185213  0.8543999  5.1826428 0.0005172
## Peter Lake-Hummingbird Lake         2.5287988  0.3685979  4.6889997 0.0086420
## Tuesday Lake-Hummingbird Lake       0.2925503 -1.9055981  2.4906986 0.9999773
## Ward Lake-Hummingbird Lake          3.6853448  0.6898445  6.6808451 0.0043115
## West Long Lake-Hummingbird Lake     0.8122467 -1.4205745  3.0450678 0.9700210
## Peter Lake-Paul Lake               -0.4897225 -1.1036180  0.1241730 0.2442990
## Tuesday Lake-Paul Lake             -2.7259711 -3.4623514 -1.9895907 0.0000000
## Ward Lake-Paul Lake                 0.6668235 -1.4972980  2.8309450 0.9895659
## West Long Lake-Paul Lake           -2.2062747 -3.0404749 -1.3720745 0.0000000
## Tuesday Lake-Peter Lake            -2.2362485 -2.9610258 -1.5114713 0.0000000
## Ward Lake-Peter Lake                1.1565460 -1.0036549  3.3167469 0.7703831
## West Long Lake-Peter Lake          -1.7165522 -2.5405279 -0.8925764 0.0000000
## Ward Lake-Tuesday Lake              3.3927945  1.1946462  5.5909429 0.0000597
## West Long Lake-Tuesday Lake         0.5196964 -0.3991749  1.4385677 0.7121762
## West Long Lake-Ward Lake           -2.8730982 -5.1059193 -0.6402770 0.0021521
```

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: Paul Lake has the same mean temperature statistically speaking as Peter Lake because when comparing their means, the p value is 0.244 which is greater than 0.05. There is no lake in this dataset that has a mean temperature that is statistically distinct from all the other lakes. I could tell because for all of them, when compared with other lakes, had a p value of greater than 0.05 at least once.

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: We can run the HSD.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?

```
summary(Lake.July.df$lakename)
```

```
## Central Long Lake     Crampton Lake    East Long Lake  Hummingbird Lake
##               119               318               968               116
##         Paul Lake        Peter Lake      Tuesday Lake         Ward Lake
##              2643              2892              1507               116
##    West Long Lake
##              1043
```

```
Lake.July.Crampton.Ward<-Lake.July.df%>%
  filter(lakename=="Crampton Lake" | lakename=="Ward Lake")
summary(Lake.July.Crampton.Ward)
```

```
##               lakename        year4          daynum          depth
##  Crampton Lake    :318   Min.   :1999   Min.   :183.0   Min.   : 0.000
##  Ward Lake        :116   1st Qu.:2004   1st Qu.:188.0   1st Qu.: 2.000
##  Central Long Lake:  0   Median :2005   Median :197.0   Median : 4.500
##  East Long Lake   :  0   Mean   :2006   Mean   :196.7   Mean   : 4.937
##  Hummingbird Lake :  0   3rd Qu.:2010   3rd Qu.:204.0   3rd Qu.: 7.000
##  Paul Lake        :  0   Max.   :2012   Max.   :211.0   Max.   :14.000
##  (Other)          :  0
##  temperature_C
##  Min.   : 5.00
##  1st Qu.: 7.40
##  Median :15.30
##  Mean   :15.11
##  3rd Qu.:22.38
##  Max.   :27.60
##
```

```
t.test(data=Lake.July.Crampton.Ward, temperature_C ~ lakename)
```

```
##
##  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: It matches my answer in part 16 in the sense that we don't reject the null hypothesis. In this case, the p-value is 0.2649. The test suggests that the temperature means of these two lakes in July are statistically equal.