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

Yin-Chia Yang

Spring 2024

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
getwd()
```

```
## [1] "/home/guest/EDA_Spring2024/EDA_forked_Spring2024"
```

```
library(tidyverse)
library(agricolae)
library(here)
here()
```

```
## [1] "/home/guest/EDA_Spring2024/EDA_forked_Spring2024"
```

```
NTL_LTER <- read.csv(
here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = TRUE)
# Set date to date format
NTL_LTER$sampledate <- as.Date(NTL_LTER$sampledate, format = "%m/%d/%y")

#2
```

```r
mytheme <- theme_classic(base_size = 12) +
   theme(
    line = element_line(
      color='magenta',
      linewidth =0.5
      ),
    legend.background = element_rect(
      color='lightgrey',
      fill = NULL
    ),
    legend.title = element_text(
      color='black'
    ),
    legend.position = "top",
    axis.text = element_text(
      color = "black"
      )
  )
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. Ha: Mean lake temperature recorded during July change 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)

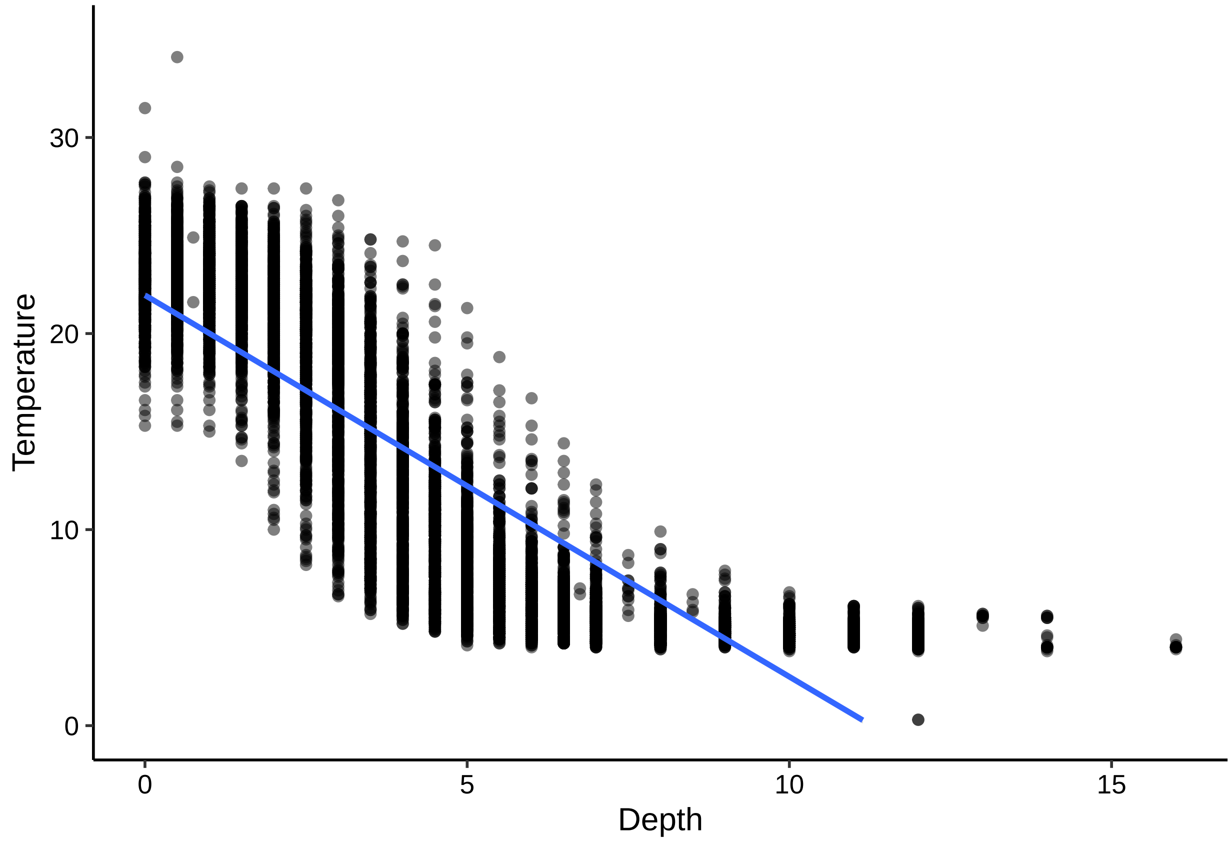
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4
NTL_LTER_Jul <- NTL_LTER %>%
  filter(month(sampledate) == 7) %>%
  select(lakename:temperature_C) %>%
  drop_na(lakename:temperature_C)

#5
library(ggplot2)

NTL_LTER_Jul_scatterplot <-
  ggplot(NTL_LTER_Jul, aes(x=depth, y=temperature_C)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = lm) +
```

```
  labs(x='Depth', y='Temperature') +
  ylim(0,35)
print(NTL_LTER_Jul_scatterplot)
```



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 scatter plot shows a roughly negative correlation between temperature and depth of lakes, though the corretlation doesn't seem to be perfectly linear.

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

```
#7
NTL_LTER_Jul.regerssion <-
  lm(NTL_LTER_Jul$temperature_C ~ NTL_LTER_Jul$depth)
summary(NTL_LTER_Jul.regerssion)
```

```
##
## Call:
## lm(formula = NTL_LTER_Jul$temperature_C ~ NTL_LTER_Jul$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_LTER_Jul$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. Also mention how much temperature is predicted to change for every 1m change in depth.

   Answer:The adjusted R-squared is 0.7387, about 74% of the variability in temperature is explained by changes in depth, based on 9726 degrees of freedom, and p-value = 2.2e-16 $<$ 0.05, which indecates the result is statistically significant. Based on the linear model, lake temperature is predicted to drop by 1.94621 degrees celsius for every 1m change in depth.

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.

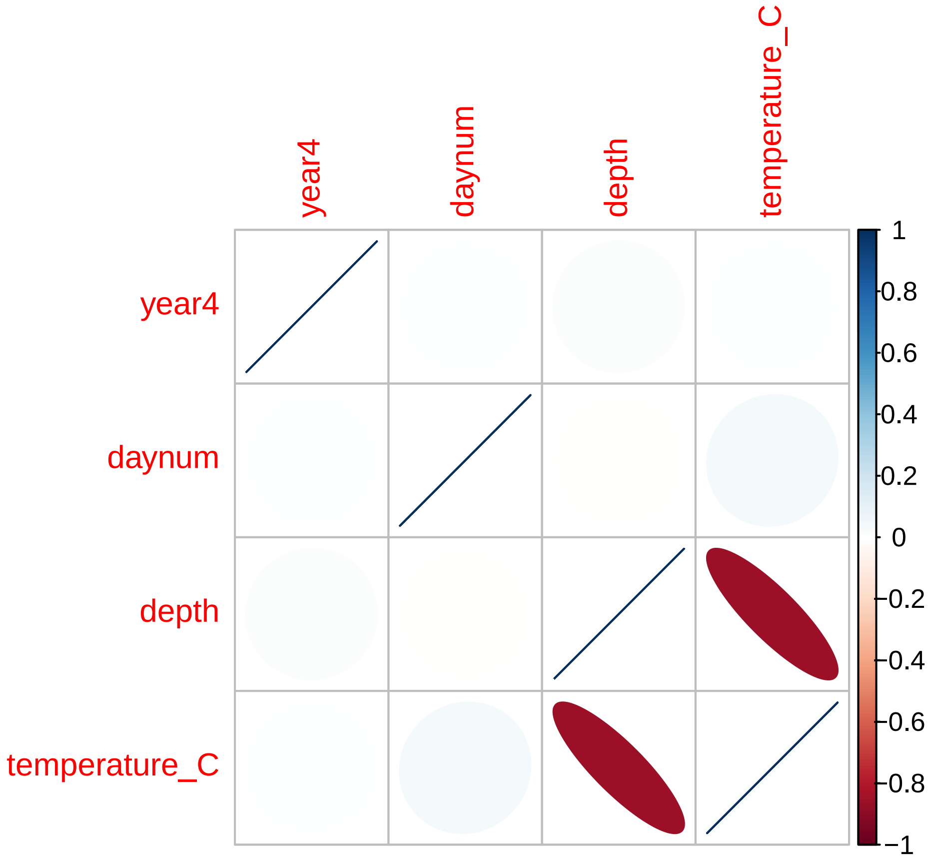
10. Run a multiple regression on the recommended set of variables.

```
#9
library(corrplot)
```

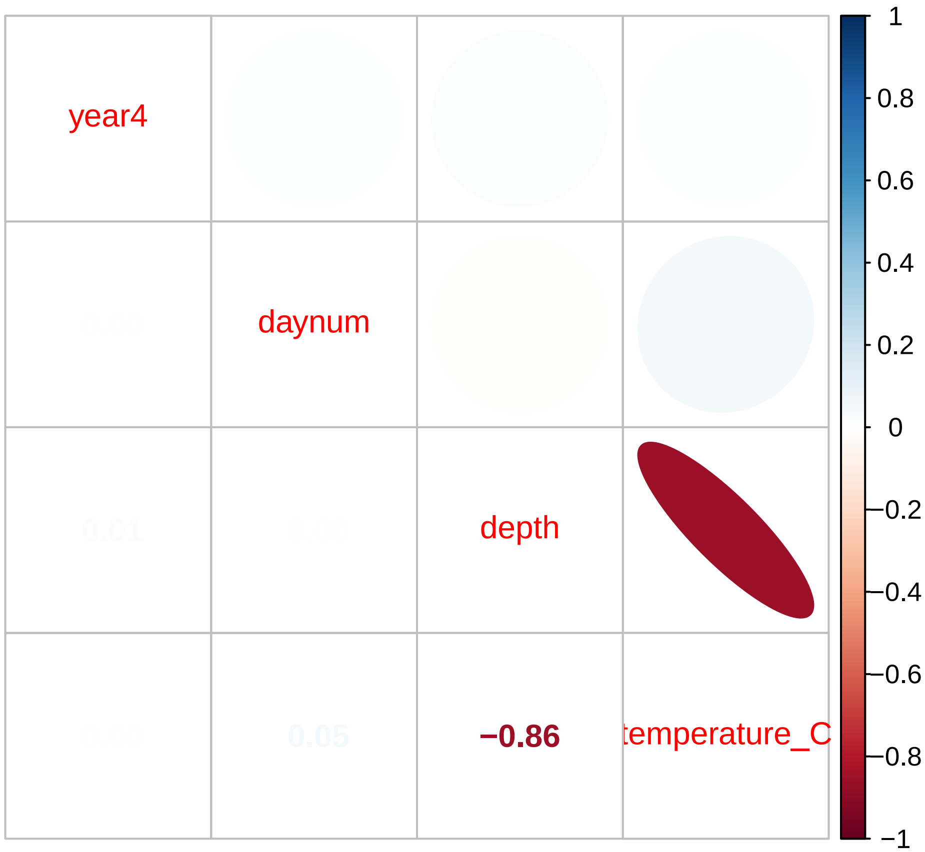
```
## corrplot 0.92 loaded
```

```
NTL_LTER_Jul.subset <- NTL_LTER_Jul %>%
  select(year4, daynum, depth, temperature_C)

corr.NTL_LTER_Jul <- cor(NTL_LTER_Jul.subset)
corrplot(corr.NTL_LTER_Jul, method = "ellipse")
```




```
corrplot.mixed(corr.NTL_LTER_Jul, upper = "ellipse")
```




```r
#10
NTL_LTER_Jul.multiregerssion <-
  lm(NTL_LTER_Jul$temperature_C ~ NTL_LTER_Jul$depth + NTL_LTER_Jul$daynum)
summary(NTL_LTER_Jul.multiregerssion)
```

```
##
## Call:
## lm(formula = NTL_LTER_Jul$temperature_C ~ NTL_LTER_Jul$depth +
##     NTL_LTER_Jul$daynum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6174 -2.9809  0.0845  2.9681 13.4406
##
## Coefficients:
##                      Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         14.088588   0.855505   16.468   <2e-16 ***
## NTL_LTER_Jul$depth  -1.946111   0.011685 -166.541   <2e-16 ***
## NTL_LTER_Jul$daynum  0.039836   0.004318    9.225   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.818 on 9725 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.741
## F-statistic: 1.391e+04 on 2 and 9725 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: I end up using 'depth' and 'daynum' as explanatory variables to predict temperature in the multipule regression model. The adjusted R-squared increased from 0.7387 to 0.741, about 74% of the variability in temperature is explained by changes in depth, based on 9725 degrees of freedom, and p-value remains the same as 2.2e-16 < 0.05, which indecates the result is statistically significant. It is slightly better than only using depth as the explanatory variable, but the improvement is not major.

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
NTL_LTER_Jul.subset2 <- NTL_LTER_Jul %>%
  group_by(lakename, sampledate, temperature_C) %>%
  summarise(temperature_C = mean(temperature_C))
```

```
## `summarise()` has grouped output by 'lakename', 'sampledate'. You can override
## using the `.groups` argument.
```

```r
NTL_LTER_Jul.anova <- aov(data = NTL_LTER_Jul.subset2, temperature_C ~ lakename)
summary(NTL_LTER_Jul.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  12361    1545   31.55 <2e-16 ***
## Residuals   8090 396256      49
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
NTL_LTER_Jul.anova2 <- lm(data = NTL_LTER_Jul.subset2, temperature_C ~ lakename)
summary(NTL_LTER_Jul.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTER_Jul.subset2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.722  -6.219  -2.691   6.881  24.009
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
```

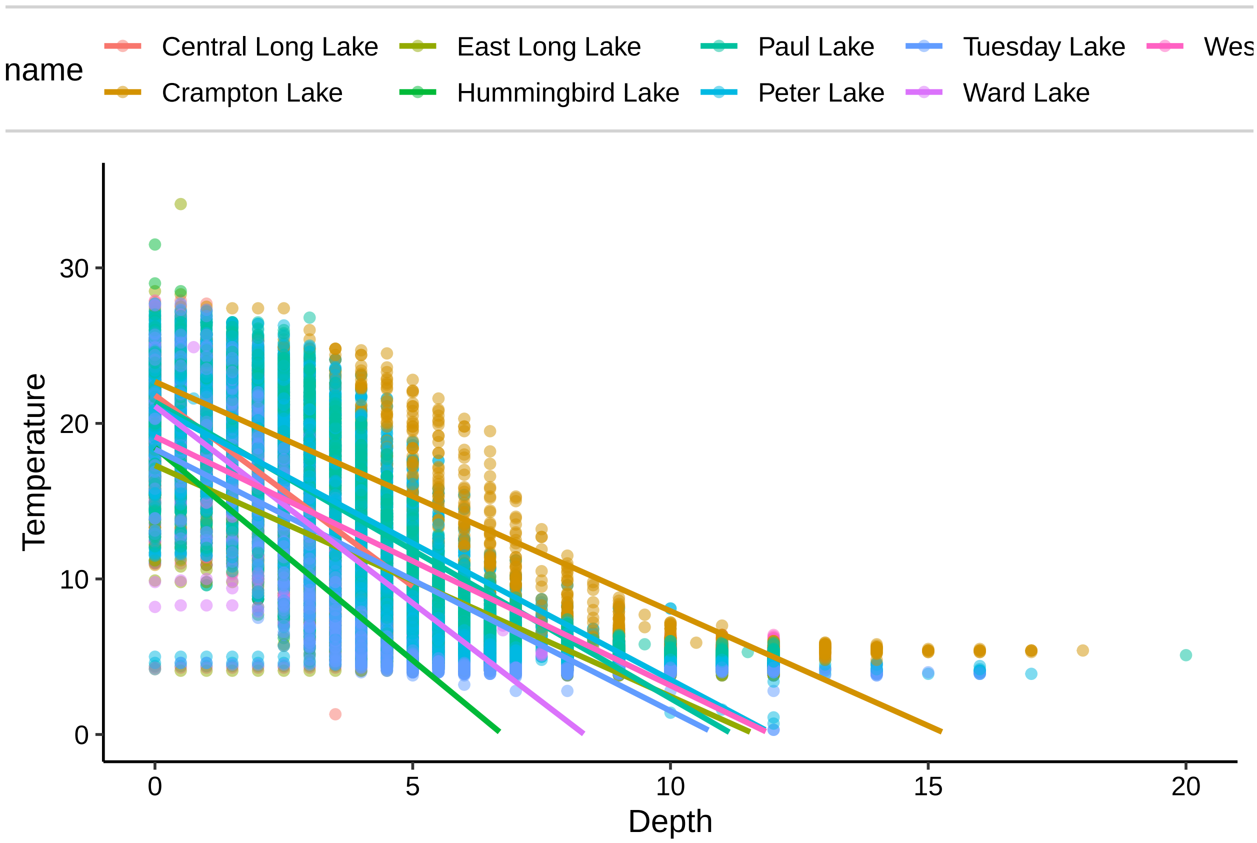
```
## (Intercept)               17.1388     0.6896  24.853  < 2e-16 ***
## lakenameCrampton Lake     -3.2021     0.8180  -3.914 9.13e-05 ***
## lakenameEast Long Lake    -7.0479     0.7317  -9.632  < 2e-16 ***
## lakenameHummingbird Lake  -6.3994     0.9617  -6.654 3.04e-11 ***
## lakenamePaul Lake         -3.9202     0.7049  -5.561 2.76e-08 ***
## lakenamePeter Lake        -4.6497     0.7044  -6.601 4.34e-11 ***
## lakenameTuesday Lake      -6.1172     0.7181  -8.518  < 2e-16 ***
## lakenameWard Lake         -2.8692     0.9554  -3.003  0.00268 **
## lakenameWest Long Lake    -5.6989     0.7319  -7.786 7.76e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.999 on 8090 degrees of freedom
## Multiple R-squared:  0.03025,    Adjusted R-squared:  0.02929
## F-statistic: 31.55 on 8 and 8090 DF,  p-value: < 2.2e-16
```

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

    Answer: Yes, with the ANOVA test, based on 8 degrees of freedom, the sum squared is 12361, F value is 31.55, with p-value $< 2e-16$, indicating not all mean temperatures of the lakes are the same, and the result is statistically significant.

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.

```r
#14.
NTL_LTER_tem.by.dep <- NTL_LTER %>%
  ggplot(aes(x=depth, y=temperature_C, color=lakename)) +
  geom_point(alpha=0.5) +
  geom_smooth(method = lm, se = FALSE) +
  labs(x='Depth', y='Temperature') +
  ylim(0,35)
print(NTL_LTER_tem.by.dep)
```



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

```
#15
NTL_LTER_Jul.groups <- HSD.test(NTL_LTER_Jul.anova, "lakename", group = TRUE)
NTL_LTER_Jul.groups$groups
```

```
##                    temperature_C groups
## Central Long Lake       17.13883      a
## Ward Lake               14.26964     ab
## Crampton Lake           13.93676      b
## Paul Lake               13.21865      b
## Peter Lake              12.48918      b
## West Long Lake          11.43993      c
## Tuesday Lake            11.02166     cd
## Hummingbird Lake        10.73945     cd
## East Long Lake          10.09095      d
```

```
NTL_LTER_Jul.subset2 <- NTL_LTER_Jul.subset2 %>%
  mutate(treatgroups = NTL_LTER_Jul.groups$groups[lakename,2])
```

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:Statistically speaking, Cramton Lake and Paul Lake has the same mean temperature as Peter Lake, and no lake has a mean temperature that is statistically distinct from all the 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: If we are only looking at two samples, we could also run a two-sample T-test to compare the mean temperatures between them.

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?

    H0:Difference in mean temperature recorded during July between Crampton Lake and Ward Lake = 0. Ha:Difference in mean temperature recorded during July between Crampton Lake and Ward Lake != 0.

```r
NTL_LTER_Jul_CramptonWard <- NTL_LTER_Jul %>%
  filter(lakename %in% c('Crampton Lake', 'Ward Lake')) %>%
  droplevels()

CramptonWard.twosample <- t.test(NTL_LTER_Jul_CramptonWard$temperature_C ~
NTL_LTER_Jul_CramptonWard$lakename)
print(CramptonWard.twosample)
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_LTER_Jul_CramptonWard$temperature_C by NTL_LTER_Jul_CramptonWard$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 r
## 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: The result from the two sample t-test shaws that there is no statistically difference in the means between Crampton Lake and Ward Lake, with the degree of freedom = 200.37, p-value = 0.26 (<0.05), so the results are statistically significant under a 95% confidence level, and we could reject the alternative hypothesis. This matches the result from Q16 with the Tukey's HSD test that the mean temperature from the two lakes are similar and could be considered equal.