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

Becca Cox

Fall 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

library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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(here)
```

```
## here() starts at /home/guest/EDE_Fall2024/EDE_Fall2024
```

```
library(lubridate)
library(dplyr)
here()
```

```
## [1] "/home/guest/EDE_Fall2024/EDE_Fall2024"
```

```
lakedata_raw <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = TRU

# Set date to date format
lakedata_raw$sampledate <- as.Date(lakedata_raw$sampledate, format = "%m/%d/%y")

#2

library(ggplot2)

#install.packages("wesanderson")

library(wesanderson)

my_theme <- theme(panel.background =
          element_rect(fill = 'lavenderblush1'),
          plot.background = element_rect(fill = "lavenderblush2"),
          axis.text = element_text(color = "darkslateblue"),
          axis.title = element_text(color = "darkslateblue"),
          plot.title = element_text(color = "darkslateblue"),
          legend.title = element_text(color = "darkslateblue"),
          legend.key = element_rect(fill = 'lavenderblush1'),
          legend.background = element_rect(fill = 'lavenderblush2'),
          legend.position = 'top')

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: The null hypothesis is that the mean lake temperature recorded in July does not change with depth across all lakes.
   Ha: The alternative hypothesis is that the mean lake temperature recorded in July does change with depth across all lakes.

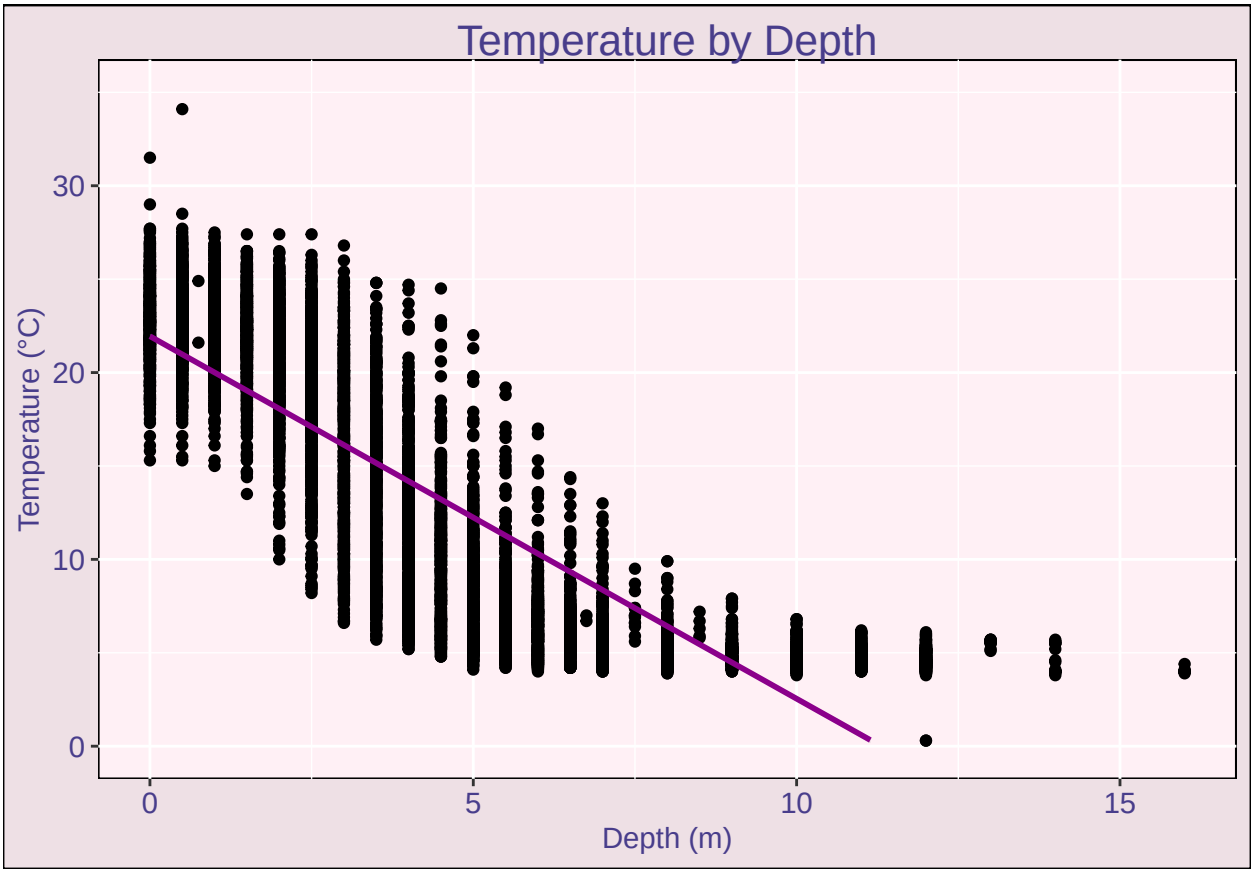
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- 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.

```
#4

lakedata_filtered <- lakedata_raw %>%
  filter(daynum > 181) %>%
  filter(daynum < 214) %>%
  select(`lakename`, `year4`, `daynum`, `depth`, `temperature_C`) %>%
  na.omit()

#5

tempbydepth <- ggplot(lakedata_filtered, aes(x = depth, y = temperature_C)) +
  geom_point() +
  scale_fill_manual(values = wes_palette("GrandBudapest2")) +
  geom_smooth(method = "lm", color = "darkmagenta") +
  ylim(0,35)+
  theme(legend.position = "none") +
  labs(title = "Temperature by Depth")+
  ylab("Temperature (°C)") +
  xlab("Depth (m)")+
  theme(plot.title = element_text(size=16))

print(tempbydepth)
```

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

## Warning: Removed 24 rows containing missing values or values outside the scale range
## (`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 suggests that there is a linear relationship between temperature and depth, temperature decreasing as depth increases. There is more variation around the trend line between depths 0 to 7m. Temperature values stabilize around a depth of 8m, at about 5°C.

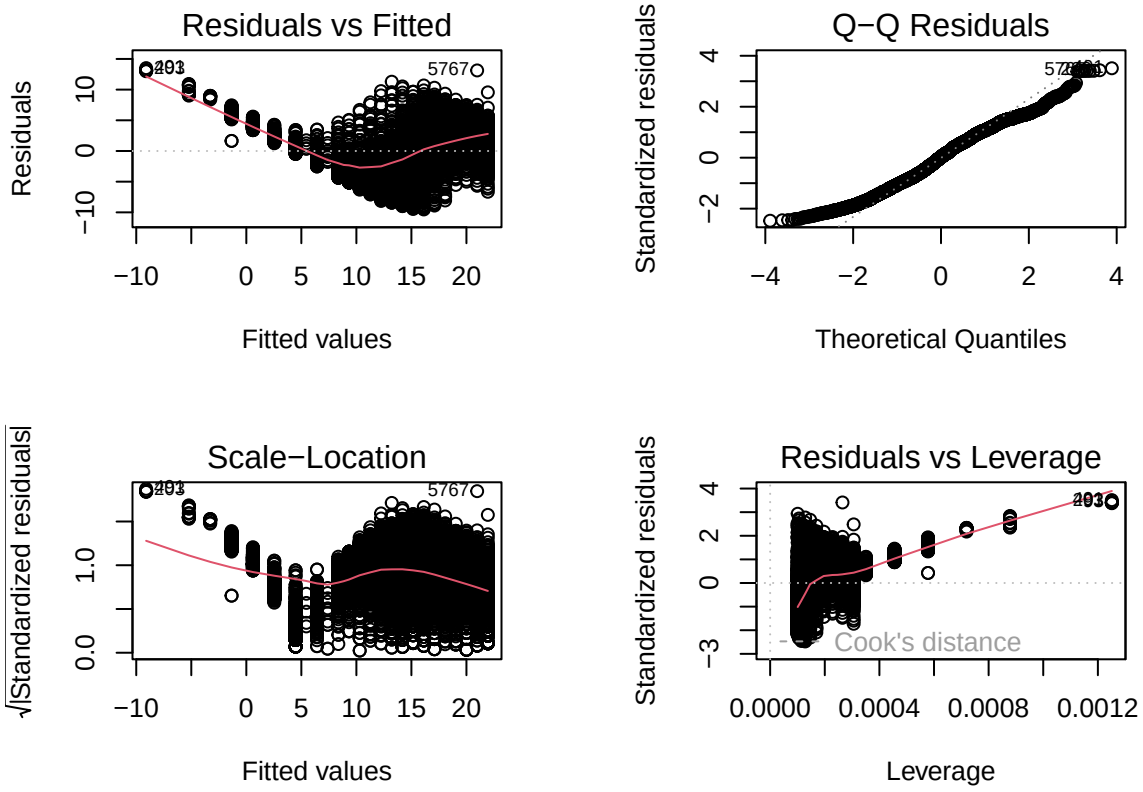
7. Perform a linear regression to test the relationship and display the results.

```
#7

tempdepth_regression <- lm(data = lakedata_filtered,
                    temperature_C ~ depth)

summary(tempdepth_regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = lakedata_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5235 -3.0362  0.0765  2.9659 13.5014
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.94464    0.06732   326.0   <2e-16 ***
## depth       -1.94038    0.01161  -167.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.849 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
```

```
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(tempdepth_regression)
```




```
par(mfrow = c(1,1))
```

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: Depth explains 73.68% of the variability in temperature, a finding that is based on 9974 degrees of freedom. This result is statistically significant, as the p value is $< 2.2e\text{-}16$, which is less than our significance level of 0.05. Temperature is predicted to decrease by 1.94038°C for every 1m increase 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

lakedata_aic <- lm(data = lakedata_filtered, temperature_C ~ depth + year4 +
              daynum)

step(lakedata_aic)
```

```
## Start:  AIC=26781.56
## temperature_C ~ depth + year4 + daynum
##
##          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 ~ depth + year4 + daynum, data = lakedata_filtered)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##   -14.33180     -1.94112      0.01386      0.04337
```

```
AIC_best <- lm(data = lakedata_filtered, temperature_C ~ depth + year4 +
              daynum)

step(AIC_best)
```

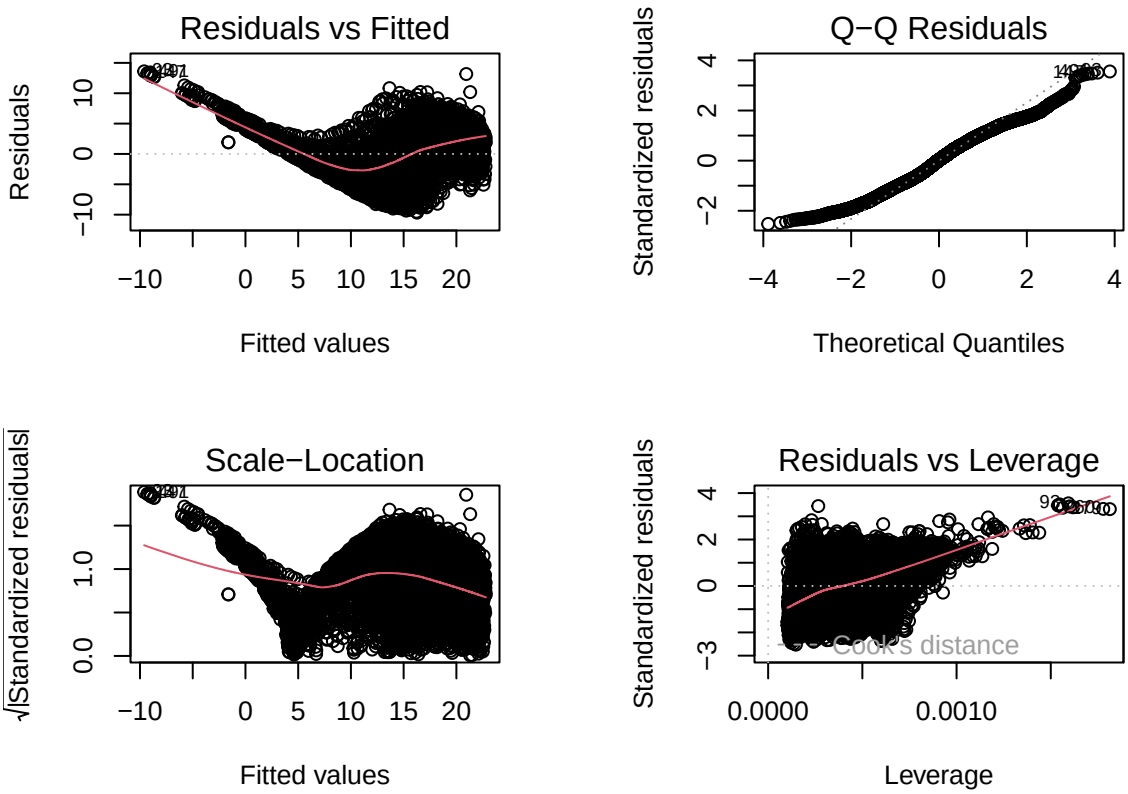
```
## Start:  AIC=26781.56
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq    RSS   AIC
## <none>                 146054 26782
## - year4   1       154 146209 26790
## - daynum  1      1582 147636 26887
## - depth   1    414049 560103 40189
```

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

```
#10

summary(AIC_best)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = lakedata_filtered)
##
## 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 .
## depth        -1.941121   0.011545 -168.135  < 2e-16 ***
## year4         0.013861   0.004274    3.243  0.00119 **
## daynum        0.043368   0.004173   10.393  < 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
```

```
par(mfrow = c(2,2), mar=c(4,4,4,4))
plot(AIC_best)
```




```
par(mfrow = c(1,1))
```

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 for predicting temperature includes depth, year, and day number. With the inclusion of all of those variables in the model, the model explains 73.99% of variation in temperature, which is a slight improvement from the model that looked only at depth, which could explain 73.68% of temperature 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

laketemp.anova <- aov(data = lakedata_filtered, temperature_C ~ lakename)
summary(laketemp.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  22407  2800.9   51.78 <2e-16 ***
## Residuals   9967 539167    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Format as lm
laketemp.anova.lm <- lm(data = lakedata_filtered, temperature_C ~ lakename)
summary(laketemp.anova.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = lakedata_filtered)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.757  -6.579  -2.680   7.686  23.813
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                17.6664     0.6501  27.175  < 2e-16 ***
## lakenameCrampton Lake      -2.1851     0.7558  -2.891 0.003847 **
## lakenameEast Long Lake     -7.3795     0.6909 -10.681  < 2e-16 ***
## lakenameHummingbird Lake   -6.5875     0.9287  -7.094 1.40e-12 ***
## lakenamePaul Lake          -3.8522     0.6655  -5.788 7.34e-09 ***
## lakenamePeter Lake         -4.3559     0.6641  -6.559 5.70e-11 ***
## lakenameTuesday Lake       -6.6094     0.6762  -9.775  < 2e-16 ***
## lakenameWard Lake          -3.2078     0.9428  -3.402 0.000671 ***
## lakenameWest Long Lake     -6.0878     0.6868  -8.864  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.355 on 9967 degrees of freedom
## Multiple R-squared:  0.0399, Adjusted R-squared:  0.03913
## F-statistic: 51.78 on 8 and 9967 DF,  p-value: < 2.2e-16
```

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

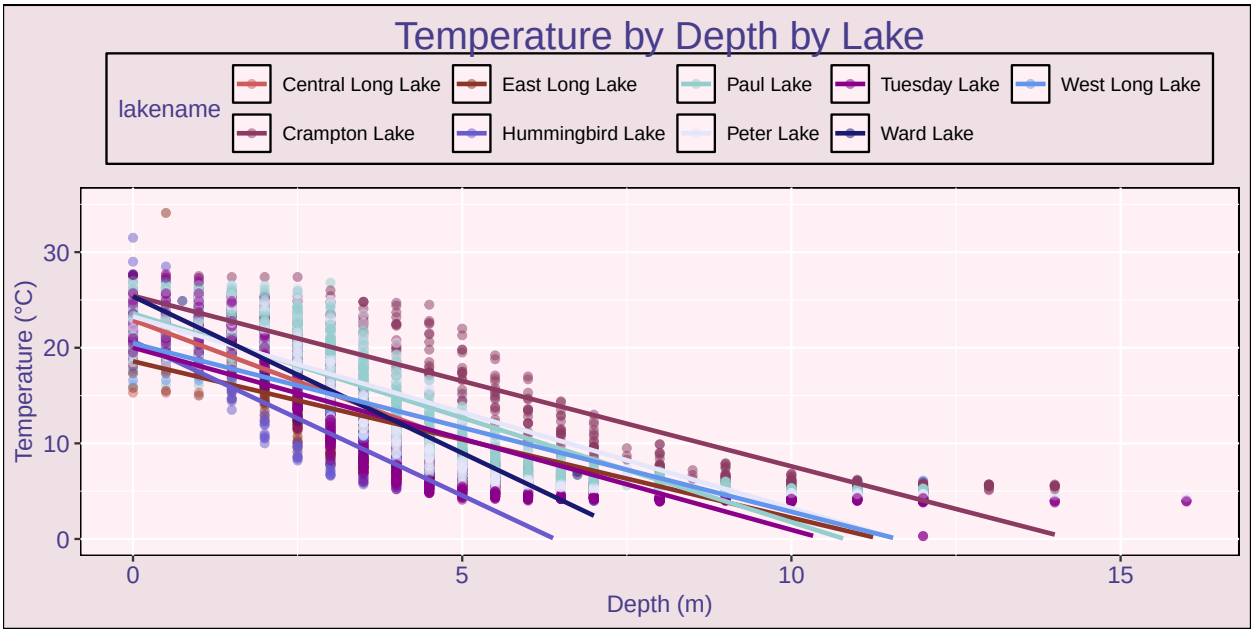
    Answer: Yes, there is a significant difference in mean temperature among the lakes. The p value from the ANOVA test is <2e-16, which is less than the significance level of 0.05. This is evident in the linear model as well: all of the lakes individually have p values of less than 0.05.

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.

tempbydepth_bylake <- lakedata_filtered %>%
  ggplot(aes(x = depth, y = temperature_C, colour = lakename)) +
  geom_point(alpha = 0.5)+
  scale_color_manual(values = c('indianred','hotpink4', 'tomato4', 'slateblue','paleturquoise3', 'laven
```

```
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0,35) +
  labs(title = "Temperature by Depth by Lake")+
  ylab("Temperature (°C)") +
  xlab("Depth (m)") +
  theme(plot.title = element_text(size=18))

print(tempbydepth_bylake)
```

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

## Warning: Removed 73 rows containing missing values or values outside the scale range
## (`geom_smooth()`).
```



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

```
#15

TukeyHSD(laketemp.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = lakedata_filtered)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.18508757 -4.5298903  0.1597151 0.0908865
## East Long Lake-Central Long Lake   -7.37946293 -9.5229855 -5.2359403 0.0000000
## Hummingbird Lake-Central Long Lake -6.58754446 -9.4686472 -3.7064417 0.0000000
## Paul Lake-Central Long Lake        -3.85215061 -5.9169602 -1.7873410 0.0000003
```

```
## Peter Lake-Central Long Lake       -4.35586017 -6.4162973 -2.2954230 0.0000000
## Tuesday Lake-Central Long Lake     -6.60943694 -8.7071991 -4.5116748 0.0000000
## Ward Lake-Central Long Lake        -3.20778556 -6.1328832 -0.2826879 0.0193258
## West Long Lake-Central Long Lake   -6.08780244 -8.2184701 -3.9571348 0.0000000
## East Long Lake-Crampton Lake       -5.19437536 -6.5934427 -3.7953080 0.0000000
## Hummingbird Lake-Crampton Lake     -4.40245689 -6.7822580 -2.0226558 0.0000004
## Paul Lake-Crampton Lake            -1.66706304 -2.9422645 -0.3918615 0.0016569
## Peter Lake-Crampton Lake           -2.17077261 -3.4388820 -0.9026632 0.0000040
## Tuesday Lake-Crampton Lake         -4.42434937 -5.7522448 -3.0964539 0.0000000
## Ward Lake-Crampton Lake            -1.02269799 -3.4555762  1.4101802 0.9304479
## West Long Lake-Crampton Lake       -3.90271487 -5.2820063 -2.5234234 0.0000000
## Hummingbird Lake-East Long Lake     0.79191847 -1.3898338  2.9736707 0.9704234
## Paul Lake-East Long Lake            3.52731232  2.6771780  4.3774466 0.0000000
## Peter Lake-East Long Lake           3.02360276  2.1841440  3.8630615 0.0000000
## Tuesday Lake-East Long Lake         0.77002599 -0.1572797  1.6973317 0.1966860
## Ward Lake-East Long Lake            4.17167737  1.9321494  6.4112054 0.0000003
## West Long Lake-East Long Lake       1.29166049  0.2921434  2.2911776 0.0020066
## Paul Lake-Hummingbird Lake          2.73539385  0.6309242  4.8398635 0.0018237
## Peter Lake-Hummingbird Lake         2.23168429  0.1315045  4.3318640 0.0273701
## Tuesday Lake-Hummingbird Lake      -0.02189248 -2.1587031  2.1149182 1.0000000
## Ward Lake-Hummingbird Lake          3.37975890  0.4265319  6.3329859 0.0115771
## West Long Lake-Hummingbird Lake     0.49974202 -1.6693818  2.6688659 0.9986055
## Peter Lake-Paul Lake               -0.50370956 -1.1148073  0.1073882 0.2050877
## Tuesday Lake-Paul Lake             -2.75728633 -3.4843663 -2.0302063 0.0000000
## Ward Lake-Paul Lake                 0.64436505 -1.5199444  2.8086745 0.9916947
## West Long Lake-Paul Lake           -2.23565183 -3.0528323 -1.4184714 0.0000000
## Tuesday Lake-Peter Lake            -2.25357677 -2.9681451 -1.5390084 0.0000000
## Ward Lake-Peter Lake                1.14807461 -1.0120638  3.3082130 0.7775157
## West Long Lake-Peter Lake          -1.73194227 -2.5380108 -0.9258737 0.0000000
## Ward Lake-Tuesday Lake              3.40165138  1.2058821  5.5974206 0.0000547
## West Long Lake-Tuesday Lake         0.52163450 -0.3755563  1.4188253 0.6798006
## West Long Lake-Ward Lake           -2.88001688 -5.1072441 -0.6527896 0.0019861
```

```
hsd_groups <- HSD.test(laketemp.anova, "lakename", group = TRUE)
hsd_groups
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   54.09519 9967 12.72153 57.81491
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9          4.38748  0.05
##
## $means
##                   temperature_C      std    r        se Min  Max   Q25   Q50
## Central Long Lake      17.66641 4.196292  128 0.6500913 8.9 26.8 14.40 18.40
## Crampton Lake          15.48132 7.347999  364 0.3855038 5.0 27.5  7.50 17.05
## East Long Lake         10.28694 6.765204  988 0.2339919 4.2 34.1  5.00  6.50
## Hummingbird Lake       11.07886 7.055590  123 0.6631729 4.0 31.5  5.25  7.70
## Paul Lake              13.81426 7.296928 2660 0.1426062 4.7 27.7  6.50 12.40
## Peter Lake             13.31055 7.671199 2930 0.1358769 4.0 27.0  5.60 11.40
## Tuesday Lake           11.05697 7.695476 1564 0.1859777 0.3 27.7  4.40  6.80
## Ward Lake              14.45862 7.409079  116 0.6828893 5.7 27.6  7.20 12.55
```

```
## West Long Lake         11.57860 6.952501 1103 0.2214581 4.0 25.7  5.40  8.00
##                       Q75
## Central Long Lake 21.000
## Crampton Lake     22.400
## East Long Lake    16.025
## Hummingbird Lake  16.400
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    18.700
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.48132     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31055      c
## West Long Lake         11.57860      d
## Hummingbird Lake       11.07886     de
## Tuesday Lake           11.05697     de
## East Long Lake         10.28694      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, statistically speaking, has the same mean temperature as Paul Lake and Ward Lake. Peter and Paul Lakes have a p value of 0.205, and Peter and Ward Lakes have a p value of 0.778; both of these are larger than the significance value of 0.05. When the lakes are assigned letter groupings, all of the lakes share a letter with at least one other lake, indicating that no one lake has a mean temperature than is statistically distinct from all of the others.

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 just looked at Peter and Paul Lakes, we could use a two-sample T-test to see if the mean temperatures are different.

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?

```
crampton_ward <- lakedata_filtered %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")
```

```r
crampton_ward_ttest <- t.test(crampton_ward$temperature_C ~ crampton_ward$lakename)
crampton_ward_ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  crampton_ward$temperature_C by crampton_ward$lakename
## t = 1.2972, df = 192.4, p-value = 0.1961
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.5323014  2.5776973
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.48132                    14.45862
```

Answer: The two-sample T-test indicates that the mean July temperatures in Crampton Lake and Ward Lake are not equal, but are not statistically different from one another. The mean temperature of Crampton Lake is 15.48°C, and the mean temperature for Ward Lake is 14.46°C. The p value is 0.1961, which is greater than the significance level of 0.05. This is the same answer that the Tukey's HSD test in part 16 provided: the mean temperatures are not statistically different. When grouped in part 16, Crampton and Ward Lakes share the "b" grouping.