Assignment 6: GLMs week 1 (t-test and ANOVA)

Masha Edmondson

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on t-tests and ANOVAs.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A06_GLMs_Week1.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 18 at 1:00 pm.

Set up your session

1. Check your working directory, load the `tidyverse`, `cowplot`, and `agricolae` packages, and import the NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv dataset.

2. Change the date column to a date format. Call up `head` of this column to verify.

```
#1
getwd()
```

```
## [1] "/Users/mashaedmondson/Desktop/Environmental_Data_Analytics_2020"
```

```
library(tidyverse)
library(agricolae)
library(cowplot)

NTL_LTER.Nutrients <- read_csv("./Data/Processed/NTL-LTER_Lake_Nutrients_PeterPaul_Processed.csv")

#2
NTL_LTER.Nutrients$sampledate <- as.Date(NTL_LTER.Nutrients$sampledate , format = "%Y-%m-%d")
head(NTL_LTER.Nutrients)
```

```
## # A tibble: 6 x 14
##   lakeid lakename year4 daynum month sampledate depth_id depth tn_ug tp_ug  nh34
##   <chr>  <chr>    <dbl>  <dbl> <dbl> <date>        <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 L      Paul La~  1991    140     5 1991-05-20        1 0       538    25    NA
## 2 L      Paul La~  1991    140     5 1991-05-20        2 0.85    285    14    NA
## 3 L      Paul La~  1991    140     5 1991-05-20        3 1.75    399    14    NA
## 4 L      Paul La~  1991    140     5 1991-05-20        4 3       453    14    NA
## 5 L      Paul La~  1991    140     5 1991-05-20        5 4       363    13    NA
## 6 L      Paul La~  1991    140     5 1991-05-20        6 6       583    37    NA
## # ... with 3 more variables: no23 <dbl>, po4 <dbl>, comments <lgl>
```

```r
NTL_LTER.Nutrients$year4 <- as.numeric(NTL_LTER.Nutrients$year4)
```

Wrangle your data

3. Wrangle your dataset so that it contains only surface depths and only the years 1993-1996, inclusive. Set month as a factor.

```r
Depths.Totals <- NTL_LTER.Nutrients %>%
 filter(year4 %in% c("1993", "1994", "1995", "1996"), depth_id == "1")

NTL_LTER.Nutrients$month <- as.factor(NTL_LTER.Nutrients$month)
Depths.Totals$month<- as.factor(Depths.Totals$month)
class(Depths.Totals$month)
```

```
## [1] "factor"
```

Analysis

Peter Lake was manipulated with additions of nitrogen and phosphorus over the years 1993-1996 in an effort to assess the impacts of eutrophication in lakes. You are tasked with finding out if nutrients are significantly higher in Peter Lake than Paul Lake, and if these potential differences in nutrients vary seasonally (use month as a factor to represent seasonality). Run two separate tests for TN and TP.

4. Which application of the GLM will you use (t-test, one-way ANOVA, two-way ANOVA with main effects, or two-way ANOVA with interaction effects)? Justify your choice.

   Answer: We would use the ANOVA GLM application to test for nitrogen and phosphorus levels between the two lakes over a three year time period. We would pick two-way anova with interaction because we have two varibales to consider: TN = Lake + month + error and TP= lake + month + error. We are are using ANOVA with interaction because we want to compare the difference in nutrient compositions between the two lakes over seasons.Though the lakes might be similar one is being dumped with nutrients, so the interaction might be able to determine that. We would not use T-test because it could only tell us one predictable variable. We would not use the one-way ANOVA because we are testing for more than one variable and we would not use the two-way ANOVA with main effects because we need to compare the interactions between the lakes.

5. Run your test for TN. Include examination of groupings and consider interaction effects, if relevant.

6. Run your test for TP. Include examination of groupings and consider interaction effects, if relevant.

```r
#5
# Format as aov
TN.anova.2way <- aov(data = Depths.Totals, tn_ug ~ month * lakename)
summary(TN.anova.2way)
```

```
##                Df  Sum Sq Mean Sq F value   Pr(>F)
## month           4  429686  107421   1.585    0.185
## lakename        1 2498451 2498451  36.855 2.47e-08 ***
## month:lakename  4  288272   72068   1.063    0.379
## Residuals      97 6575834   67792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 23 observations deleted due to missingness
```

```r
# This summary tells us that the only significant difference is between Peter and Paul lake. Therefore,

#6
```

```r
# Format as aov
TP.anova.2way <- aov(data = Depths.Totals, tp_ug ~ month * lakename)
summary(TP.anova.2way) #sig difference between lake name and month and we want to inperpret interaction
```

```
##                  Df Sum Sq Mean Sq F value Pr(>F)
## month             4    671     168   1.623 0.1730
## lakename          1  10370   10370 100.283 <2e-16 ***
## month:lakename    4   1014     254   2.452 0.0496 *
## Residuals       119  12305     103
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

```r
# Format as lm
TP.anova.2way2 <- lm(data = Depths.Totals, tp_ug ~ month * lakename)
summary(TP.anova.2way2)
```

```
##
## Call:
## lm(formula = tp_ug ~ month * lakename, data = Depths.Totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.384  -4.473  -0.693   1.939  32.489
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                11.4740     4.1514   2.764  0.00662 **
## month6                     -0.9179     4.8288  -0.190  0.84957
## month7                     -1.7271     4.7936  -0.360  0.71927
## month8                     -2.0872     4.7936  -0.435  0.66405
## month9                     -0.7380     6.1575  -0.120  0.90480
## lakenamePeter Lake          4.3136     5.6574   0.762  0.44729
## month6:lakenamePeter Lake  13.4882     6.6207   2.037  0.04384 *
## month7:lakenamePeter Lake  20.3440     6.6207   3.073  0.00263 **
## month8:lakenamePeter Lake  12.7937     6.5722   1.947  0.05394 .
## month9:lakenamePeter Lake  11.1697     8.8622   1.260  0.21000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.17 on 119 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.4949, Adjusted R-squared:  0.4567
## F-statistic: 12.95 on 9 and 119 DF,  p-value: 3.24e-14
```

```r
TukeyHSD(TP.anova.2way)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = tp_ug ~ month * lakename, data = Depths.Totals)
##
## $month
##           diff        lwr       upr     p adj
## 6-5  5.9146220  -3.234390 15.063634 0.3837749
```

```
## 7-5  7.9267363  -1.222276 17.075748 0.1224572
## 8-5  4.3748753  -4.706921 13.456671 0.6703911
## 9-5  3.8207521  -8.393804 16.035308 0.9085595
## 7-6  2.0121143  -4.721376  8.745605 0.9215444
## 8-6 -1.5397467  -8.181621  5.102128 0.9677800
## 9-6 -2.0938698 -12.621493  8.433754 0.9816312
## 8-7 -3.5518610 -10.193735  3.090013 0.5765788
## 9-7 -4.1059841 -14.633608  6.421639 0.8162959
## 9-8 -0.5541231 -11.023385  9.915139 0.9998946
##
## $lakename
##                          diff      lwr      upr p adj
## Peter Lake-Paul Lake 17.91381 14.36807 21.45955     0
##
## $`month:lakename`
##                                diff         lwr       upr     p adj
## 6:Paul Lake-5:Paul Lake   -0.9178824 -16.4886641 14.652899 1.0000000
## 7:Paul Lake-5:Paul Lake   -1.7271111 -17.1846493 13.730427 0.9999981
## 8:Paul Lake-5:Paul Lake   -2.0872222 -17.5447604 13.370316 0.9999902
## 9:Paul Lake-5:Paul Lake   -0.7380000 -20.5935673 19.117567 1.0000000
## 5:Peter Lake-5:Paul Lake   4.3135714 -13.9293175 22.556460 0.9989515
## 6:Peter Lake-5:Paul Lake  16.8838889   1.4263507 32.341427 0.0206973
## 7:Peter Lake-5:Paul Lake  22.9304706   7.3596889 38.501252 0.0002415
## 8:Peter Lake-5:Paul Lake  15.0200000  -0.3355071 30.375507 0.0607728
## 9:Peter Lake-5:Paul Lake  14.7452500  -6.4208558 35.911356 0.4316694
## 7:Paul Lake-6:Paul Lake   -0.8092288 -11.8989312 10.280474 1.0000000
## 8:Paul Lake-6:Paul Lake   -1.1693399 -12.2590423  9.920363 0.9999989
## 9:Paul Lake-6:Paul Lake    0.1798824 -16.5021309 16.861896 1.0000000
## 5:Peter Lake-6:Paul Lake   5.2314538  -9.4943403 19.957248 0.9787107
## 6:Peter Lake-6:Paul Lake  17.8017712   6.7120688 28.891474 0.0000401
## 7:Peter Lake-6:Paul Lake  23.8483529  12.6013419 35.095364 0.0000000
## 8:Peter Lake-6:Paul Lake  15.9378824   4.9908457 26.884919 0.0003006
## 9:Peter Lake-6:Paul Lake  15.6631324  -2.5591082 33.885373 0.1584032
## 8:Paul Lake-7:Paul Lake   -0.3601111 -11.2902412 10.570019 1.0000000
## 9:Paul Lake-7:Paul Lake    0.9891111 -15.5872518 17.565474 1.0000000
## 5:Peter Lake-7:Paul Lake   6.0406825  -8.5653181 20.646683 0.9437275
## 6:Peter Lake-7:Paul Lake  18.6110000   7.6808700 29.541130 0.0000101
## 7:Peter Lake-7:Paul Lake  24.6575817  13.5678793 35.747284 0.0000000
## 8:Peter Lake-7:Paul Lake  16.7471111   5.9617574 27.532465 0.0000827
## 9:Peter Lake-7:Paul Lake  16.4723611  -1.6532090 34.597931 0.1087387
## 9:Paul Lake-8:Paul Lake    1.3492222 -15.2271407 17.925585 0.9999999
## 5:Peter Lake-8:Paul Lake   6.4007937  -8.2052070 21.006794 0.9208652
## 6:Peter Lake-8:Paul Lake  18.9711111   8.0409811 29.901241 0.0000062
## 7:Peter Lake-8:Paul Lake  25.0176928  13.9279904 36.107395 0.0000000
## 8:Peter Lake-8:Paul Lake  17.1072222   6.3218685 27.892576 0.0000523
## 9:Peter Lake-8:Paul Lake  16.8324722  -1.2930979 34.958042 0.0926020
## 5:Peter Lake-9:Paul Lake   5.0515714 -14.1485150 24.251658 0.9975850
## 6:Peter Lake-9:Paul Lake  17.6218889   1.0455259 34.198252 0.0276305
## 7:Peter Lake-9:Paul Lake  23.6684706   6.9864574 40.350484 0.0004851
## 8:Peter Lake-9:Paul Lake  15.7580000  -0.7232597 32.239260 0.0735733
## 9:Peter Lake-9:Paul Lake  15.4832500  -6.5132124 37.479712 0.4163366
## 6:Peter Lake-5:Peter Lake 12.5703175  -2.0356832 27.176318 0.1571717
## 7:Peter Lake-5:Peter Lake 18.6168992   3.8911050 33.342693 0.0032014
## 8:Peter Lake-5:Peter Lake 10.7064286  -3.7915495 25.204407 0.3464892
```

```
## 9:Peter Lake-5:Peter Lake 10.4316786 -10.1207861 30.984143 0.8273658
## 7:Peter Lake-6:Peter Lake  6.0465817  -5.0431207 17.136284 0.7595330
## 8:Peter Lake-6:Peter Lake -1.8638889 -12.6492426  8.921465 0.9999197
## 9:Peter Lake-6:Peter Lake -2.1386389 -20.2642090 15.986931 0.9999970
## 8:Peter Lake-7:Peter Lake -7.9104706 -18.8575073  3.036566 0.3778093
## 9:Peter Lake-7:Peter Lake -8.1852206 -26.4074611 10.037020 0.9089776
## 9:Peter Lake-8:Peter Lake -0.2747500 -18.3133864 17.763886 1.0000000
```

```
TP_interaction.lake.month <-with(Depths.Totals, interaction(month, lakename))
TP.interaction.anova <- aov(data = Depths.Totals, tp_ug ~ TP_interaction.lake.month)

TP_interaction_groups <-HSD.test(TP.interaction.anova, "TP_interaction.lake.month", group = TRUE)
TP_interaction_groups
```

```
## $statistics
##    MSerror  Df     Mean      CV
##   103.4055 119 19.07347 53.3141
##
## $parameters
##    test                    name.t ntr StudentizedRange alpha
##   Tukey TP_interaction.lake.month  10         4.560262  0.05
##
## $means
##                  tp_ug       std  r    Min    Max     Q25     Q50      Q75
## 5.Paul Lake  11.474000  3.928545  6  7.001 17.090  8.1395 11.8885 13.53675
## 5.Peter Lake 15.787571  2.719954  7 10.887 18.922 14.8915 15.5730 17.67400
## 6.Paul Lake  10.556118  4.416821 17  1.222 16.697  7.4430 10.6050 13.94600
## 6.Peter Lake 28.357889 15.588507 18 10.974 53.388 14.7790 24.6840 41.13000
## 7.Paul Lake   9.746889  3.525120 18  4.501 21.763  7.8065  9.1555 10.65700
## 7.Peter Lake 34.404471 18.285568 17 19.149 66.893 21.6640 24.2070 50.54900
## 8.Paul Lake   9.386778  1.478062 18  5.879 11.542  8.4495  9.6090 10.45050
## 8.Peter Lake 26.494000  9.829596 19 14.551 49.757 21.2425 23.2250 27.99350
## 9.Paul Lake  10.736000  3.615978  5  6.592 16.281  8.9440 10.1920 11.67100
## 9.Peter Lake 26.219250 10.814803  4 16.281 41.145 19.6845 23.7255 30.26025
##
## $comparison
## NULL
##
## $groups
##                  tp_ug groups
## 7.Peter Lake 34.404471      a
## 6.Peter Lake 28.357889     ab
## 8.Peter Lake 26.494000    abc
## 9.Peter Lake 26.219250   abcd
## 5.Peter Lake 15.787571    bcd
## 5.Paul Lake  11.474000     cd
## 9.Paul Lake  10.736000     cd
## 6.Paul Lake  10.556118      d
## 7.Paul Lake   9.746889      d
## 8.Paul Lake   9.386778      d
##
## attr(,"class")
## [1] "group"
```

7. Create two plots, with TN (plot 1) or TP (plot 2) as the response variable and month and lake as the

predictor variables. Hint: you may use some of the code you used for your visualization assignment. Assign groupings with letters, as determined from your tests. Adjust your axes, aesthetics, and color palettes in accordance with best data visualization practices.

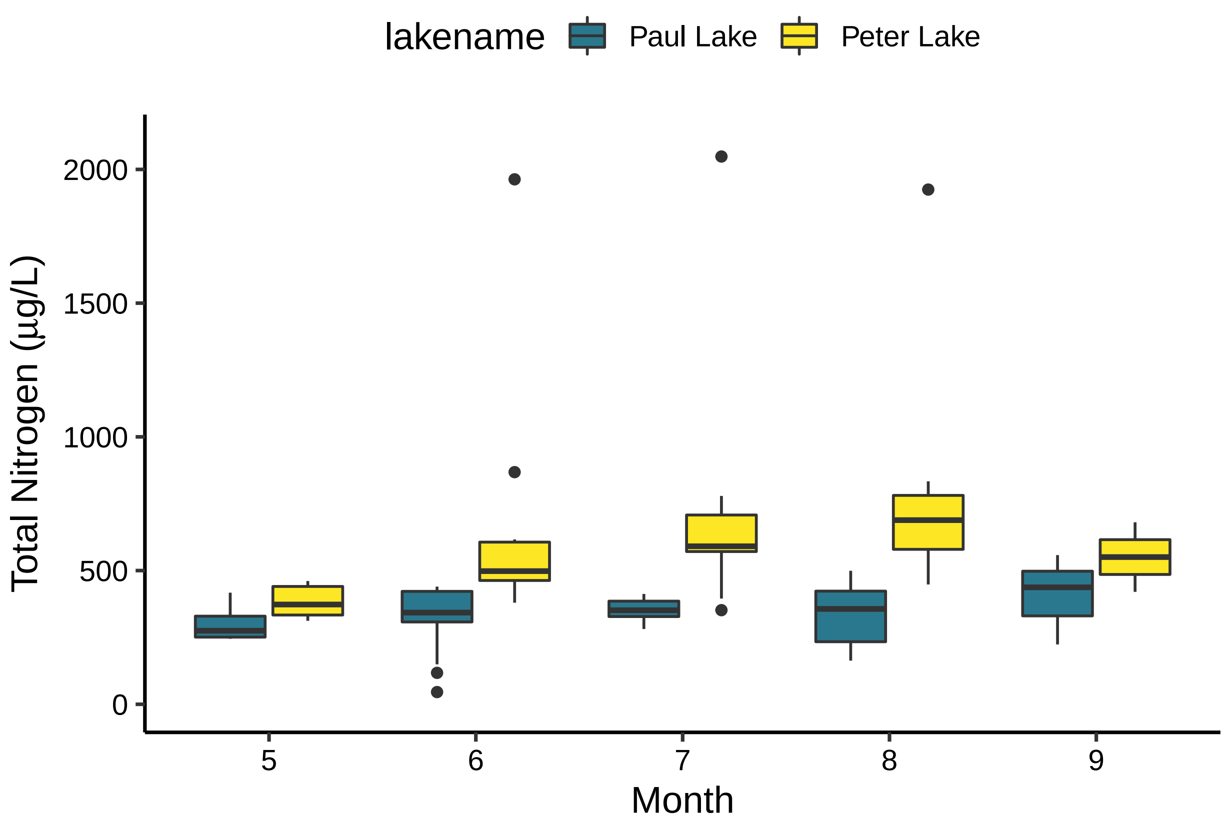
8. Combine your plots with cowplot, with a common legend at the top and the two graphs stacked vertically. Your x axes should be formatted with the same breaks, such that you can remove the title and text of the top legend and retain just the bottom legend.

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

theme_set(settheme)

#7 - change the color so it is not the default and set a theme for the graph
TN.anova.plot <- ggplot(Depths.Totals, aes(y = tn_ug, x = month, fill= lakename)) +
  geom_boxplot()+
  scale_fill_viridis_d(begin = 0.4, end = 1.0)+
  labs(x = "Month", y = expression(paste("Total Nitrogen (", mu, "g/L)"))) +
  ylim(0, 2100)
print(TN.anova.plot)
```

```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).
```

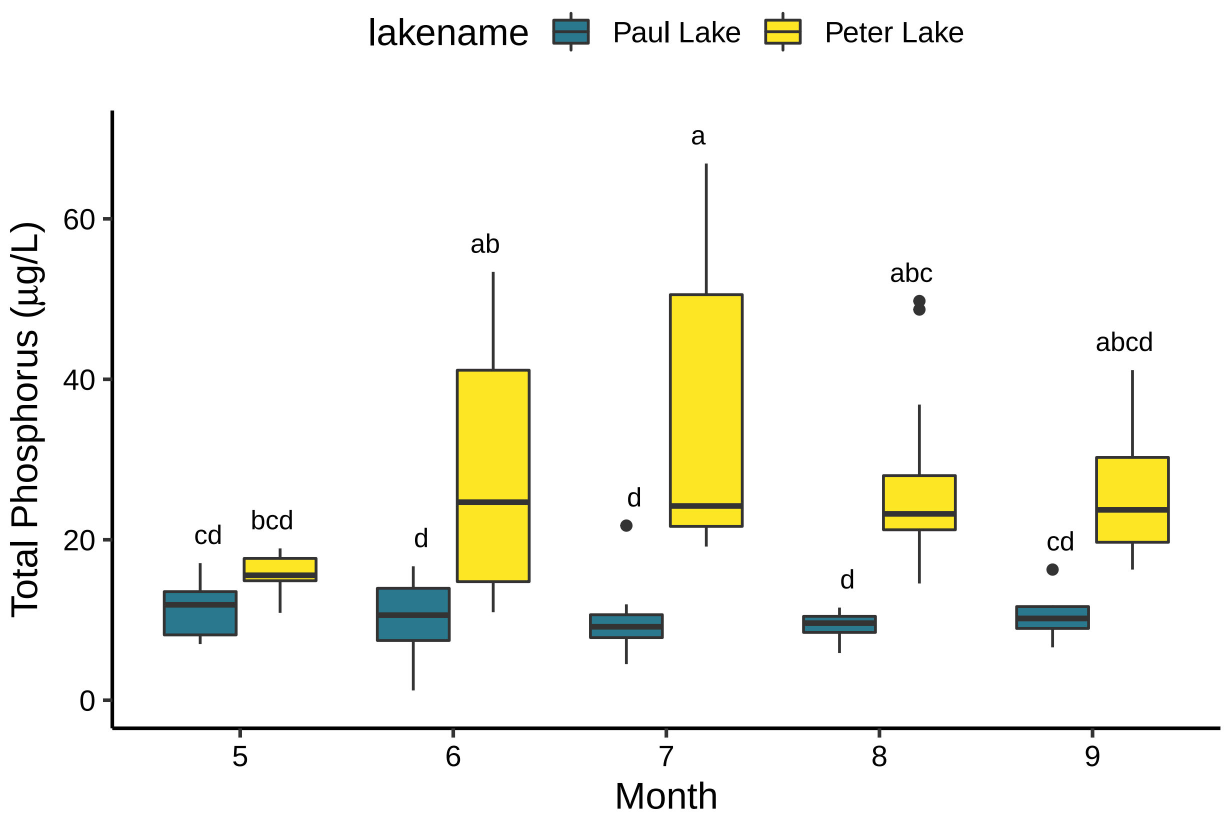



```
TP.anova.plot <- ggplot(Depths.Totals, aes(y = tp_ug, x = month, fill = lakename)) +
  geom_boxplot()+
  scale_fill_viridis_d(begin = 0.4, end = 1.0)+
  labs(x = "Month", y = expression(paste("Total Phosphorus (", mu, "g/L)")))+
```

```r
  stat_summary(geom = "text", fun.y = max, vjust = -1, size = 3.5,
               label = c("cd", "bcd", "d", "ab", "d", "a", "d", "abc", "cd", "abcd"),
               show.legend= FALSE,
               position = position_dodge2(0.6))+
  ylim(0, 70)
print(TP.anova.plot)
```

```
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_summary).
```

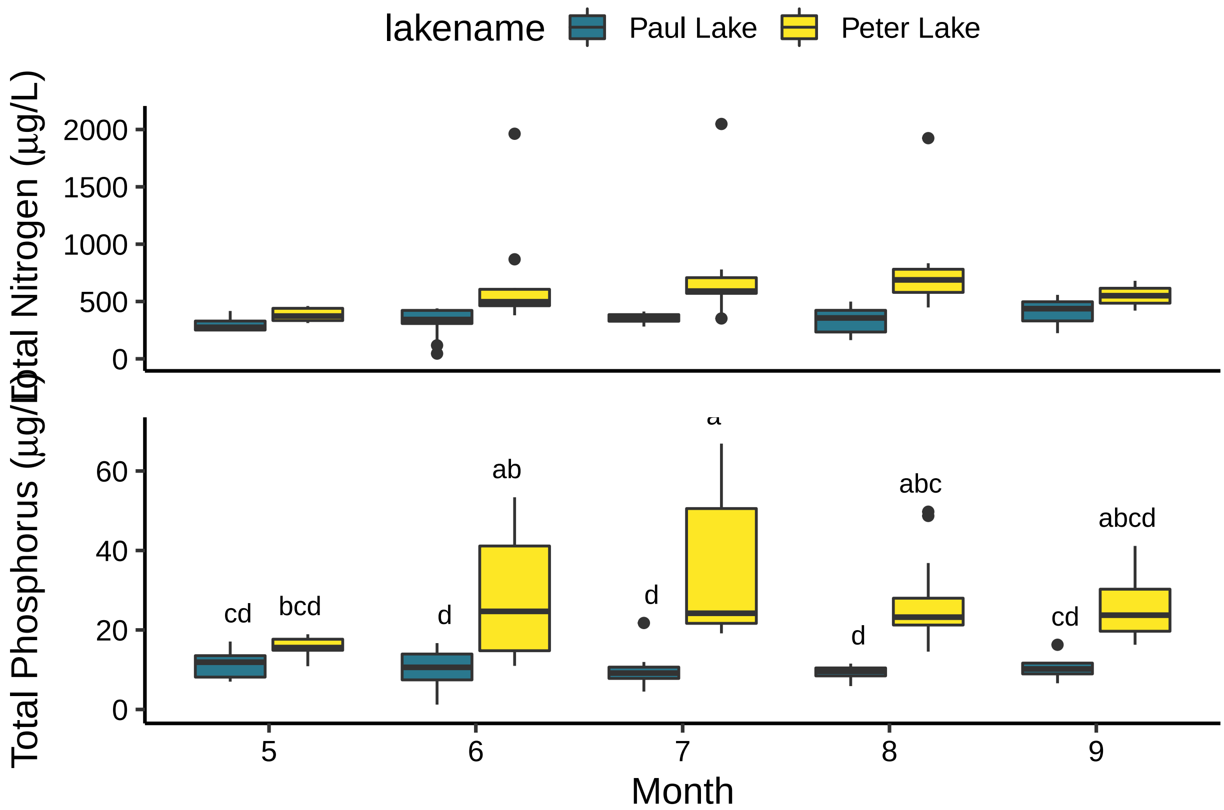



```r
#8
plot_grid(TN.anova.plot + theme(axis.text.x = element_blank(),
                                axis.ticks.x= element_blank(),
                                axis.title.x = element_blank()),
          TP.anova.plot + theme(legend.position = "none"),
                                 align = "v", nrow = 2, ncol=1)
```

```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_boxplot).

## Warning: Removed 1 rows containing non-finite values (stat_summary).
```



lakename
Paul Lake
Peter Lake
Total Nitrogen (µg/L)
2000
1500
1000
500
0
Total Phosphorus (µg/L)
60
40
20
0
a
ab
abc
abcd
cd
bcd
d
d
d
cd
5
6
7
8
9
Month