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

Cristiana Falvo

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/cristiana/Documents/Duke/DataAnalytics/Environmental_Data_Analytics_2020/Assignments"
```

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(cowplot)
library(agricolae)
Nutrients <- read.csv("/Users/cristiana/Documents/Duke/DataAnalytics/Environmental_Data_Analytics_2020/

#2
head(Nutrients)
```

```
##   lakeid  lakename year4 daynum month sampledate depth_id depth tn_ug tp_ug
## 1      L Paul Lake  1991    140     5 1991-05-20        1  0.00   538    25
## 2      L Paul Lake  1991    140     5 1991-05-20        2  0.85   285    14
## 3      L Paul Lake  1991    140     5 1991-05-20        3  1.75   399    14
## 4      L Paul Lake  1991    140     5 1991-05-20        4  3.00   453    14
## 5      L Paul Lake  1991    140     5 1991-05-20        5  4.00   363    13
## 6      L Paul Lake  1991    140     5 1991-05-20        6  6.00   583    37
##   nh34 no23 po4 comments
## 1   NA   NA  NA       NA
## 2   NA   NA  NA       NA
```

```
## 3   NA   NA  NA       NA
## 4   NA   NA  NA       NA
## 5   NA   NA  NA       NA
## 6   NA   NA  NA       NA
```

```
Nutrients$sampledate <- as.Date(Nutrients$sampledate, format = "%Y-%m-%d") # not working
head(Nutrients)
```

```
##   lakeid  lakename year4 daynum month sampledate depth_id depth tn_ug tp_ug
## 1      L Paul Lake  1991    140     5 1991-05-20        1  0.00   538    25
## 2      L Paul Lake  1991    140     5 1991-05-20        2  0.85   285    14
## 3      L Paul Lake  1991    140     5 1991-05-20        3  1.75   399    14
## 4      L Paul Lake  1991    140     5 1991-05-20        4  3.00   453    14
## 5      L Paul Lake  1991    140     5 1991-05-20        5  4.00   363    13
## 6      L Paul Lake  1991    140     5 1991-05-20        6  6.00   583    37
##   nh34 no23 po4 comments
## 1   NA   NA  NA       NA
## 2   NA   NA  NA       NA
## 3   NA   NA  NA       NA
## 4   NA   NA  NA       NA
## 5   NA   NA  NA       NA
## 6   NA   NA  NA       NA
```

```
str(Nutrients)
```

```
## 'data.frame':    2406 obs. of  14 variables:
##  $ lakeid    : Factor w/ 2 levels "L","R": 1 1 1 1 1 1 2 2 2 2 ...
##  $ lakename  : Factor w/ 2 levels "Paul Lake","Peter Lake": 1 1 1 1 1 1 2 2 2 2 ...
##  $ year4     : int  1991 1991 1991 1991 1991 1991 1991 1991 1991 1991 ...
##  $ daynum    : int  140 140 140 140 140 140 140 140 140 140 ...
##  $ month     : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ sampledate: Date, format: "1991-05-20" "1991-05-20" ...
##  $ depth_id  : int  1 2 3 4 5 6 1 2 3 4 ...
##  $ depth     : num  0 0.85 1.75 3 4 6 0 1 2.25 3.5 ...
##  $ tn_ug     : num  538 285 399 453 363 583 352 356 364 582 ...
##  $ tp_ug     : num  25 14 14 14 13 37 11 15 28 14 ...
##  $ nh34      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ no23      : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ po4       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ comments  : logi  NA NA NA NA NA NA ...
```

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.

```
subset <- filter(Nutrients, depth == 0, year4 >= 1993 & year4 <= 1996)

subset$month <- as.factor(subset$month)

class(Nutrients$year4)
```

```
## [1] "integer"
```

```
class(Nutrients$month)
```

```
## [1] "integer"
```

```r
class(subset$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: t-test -two independent samples test (peter and paul lakes are two independent groups we want to compare) -two-sided test (seeing if there's difference, nondirectional meaning neither is the standard)

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.

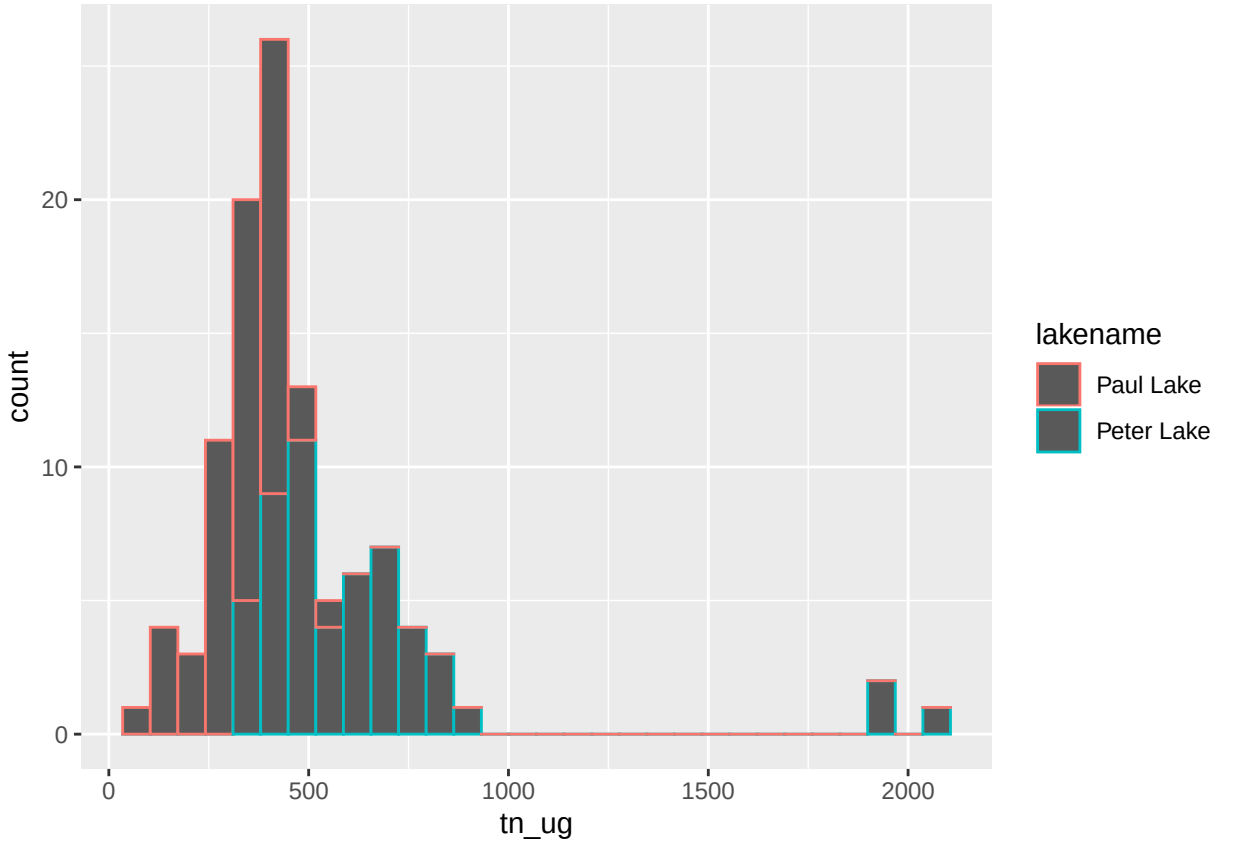
```r
#5
shapiro.test(subset$tn_ug)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset$tn_ug
## W = 0.67197, p-value = 3.969e-14
```

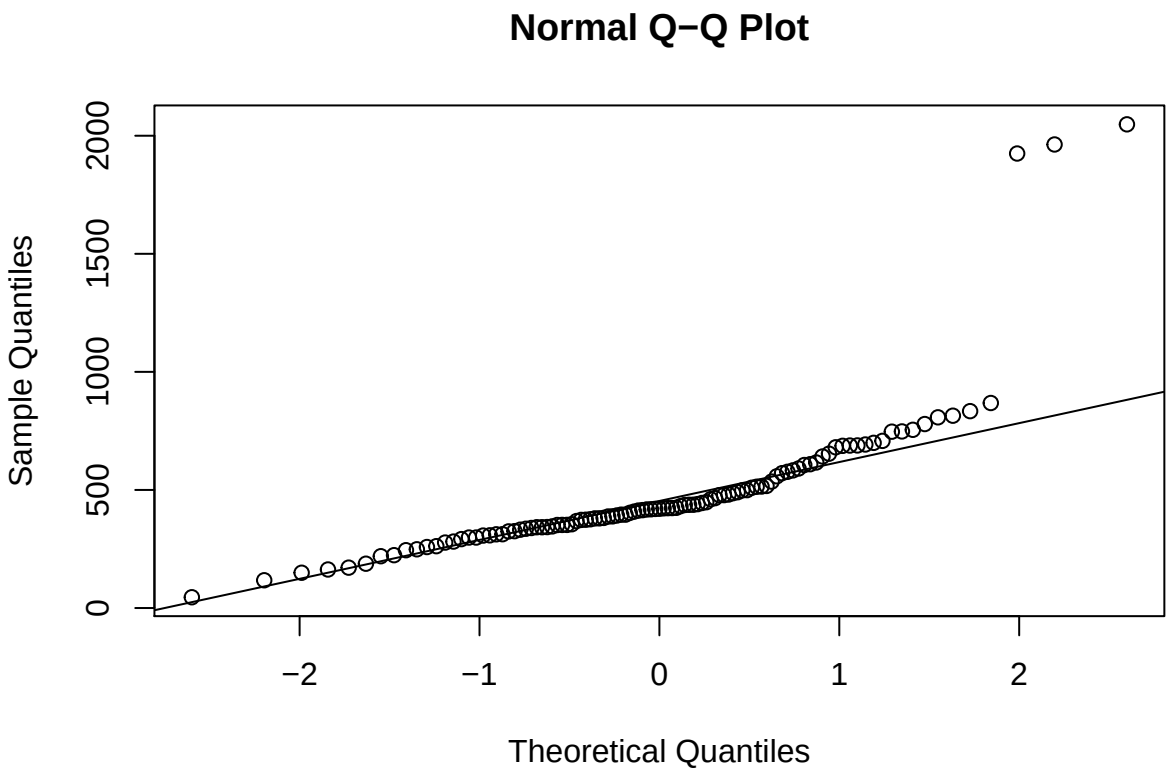
```r
ggplot(subset, aes(x = tn_ug, color = lakename)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
qqnorm(subset$tn_ug); qqline(subset$tn_ug)
```




```
O3.onesample <- t.test(subset$tn_ug, mu = 50, alternative = "two.sided")
O3.onesample
```

```
##
##  One Sample t-test
##
## data:  subset$tn_ug
## t = 14.886, df = 106, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  429.1530 545.6624
## sample estimates:
## mean of x
##  487.4077
```

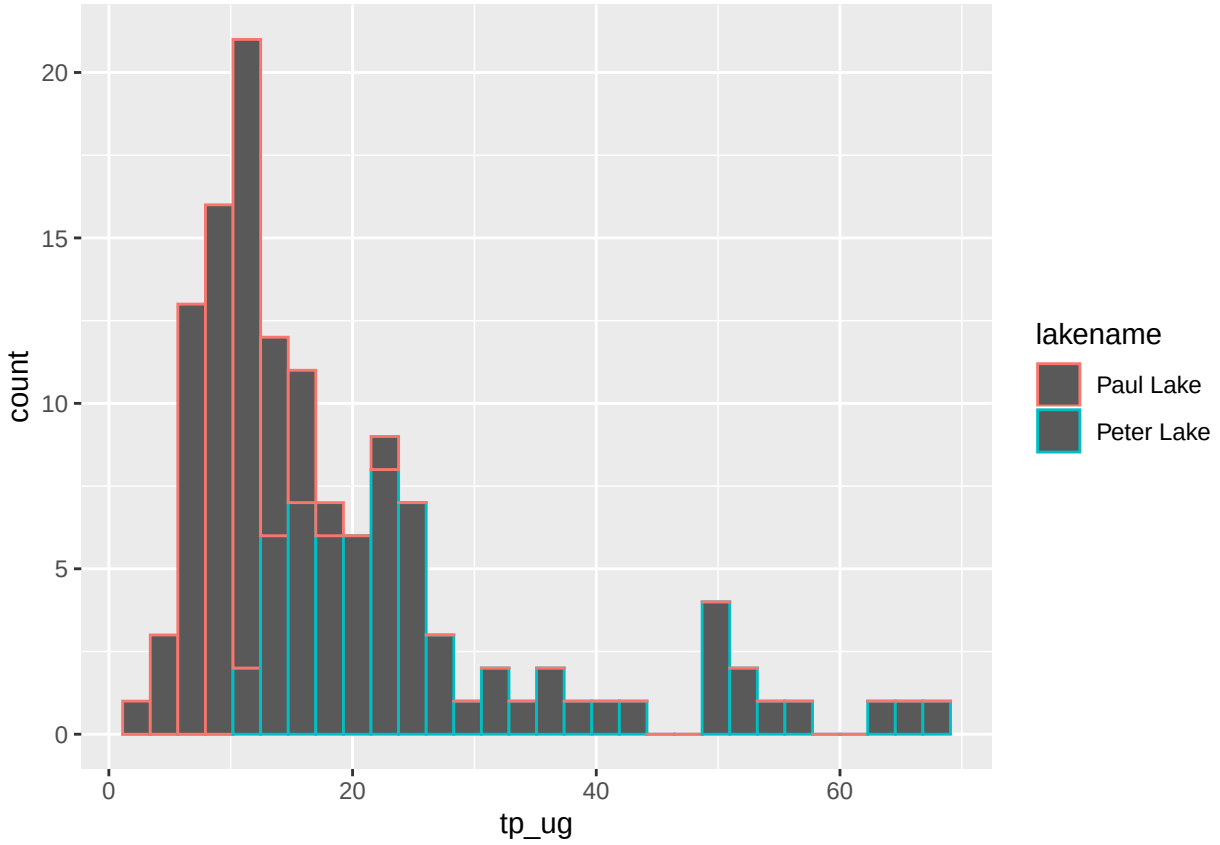
```
#6
shapiro.test(subset$tp_ug)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  subset$tp_ug
## W = 0.80421, p-value = 7.857e-12
```

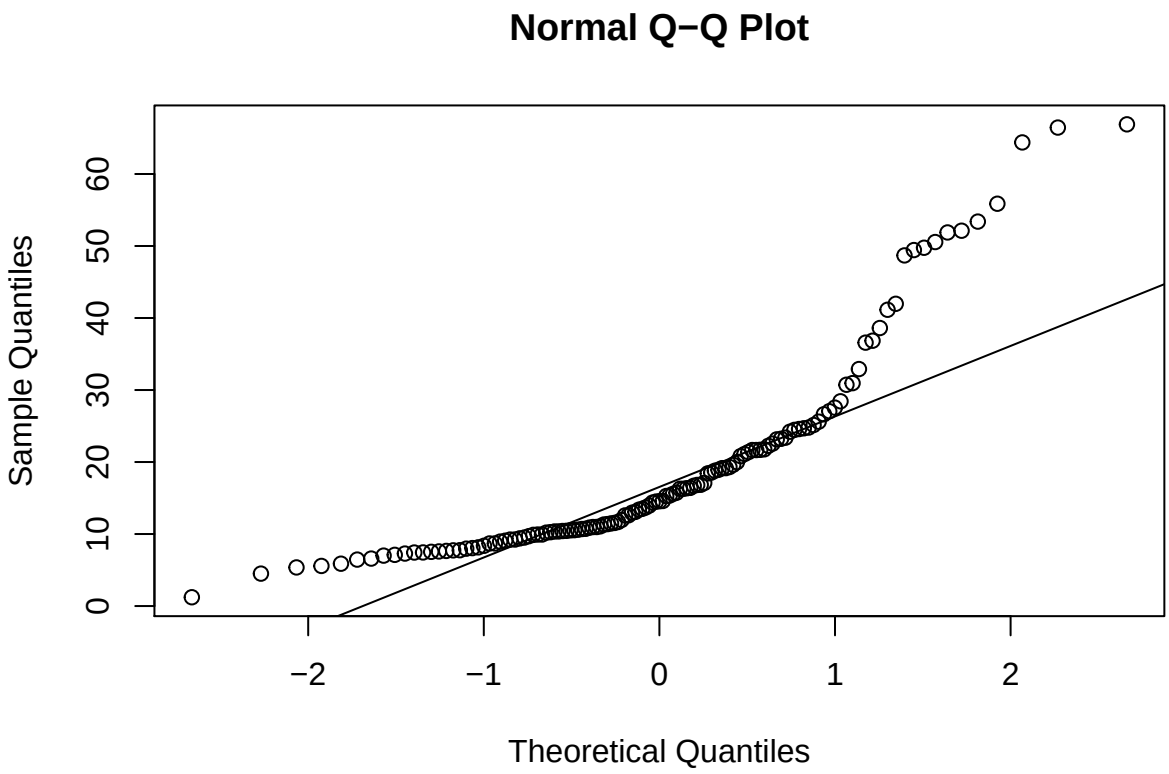
```
ggplot(subset, aes(x = tp_ug, color = lakename)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

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




```
qqnorm(subset$tp_ug); qqline(subset$tp_ug)
```




```
O3.onesample <- t.test(subset$tp_ug, mu = 50, alternative = "two.sided")
O3.onesample
```

```
##
##  One Sample t-test
##
## data:  subset$tp_ug
## t = -25.462, df = 128, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  16.67012 21.47683
## sample estimates:
## mean of x
##  19.07347
```

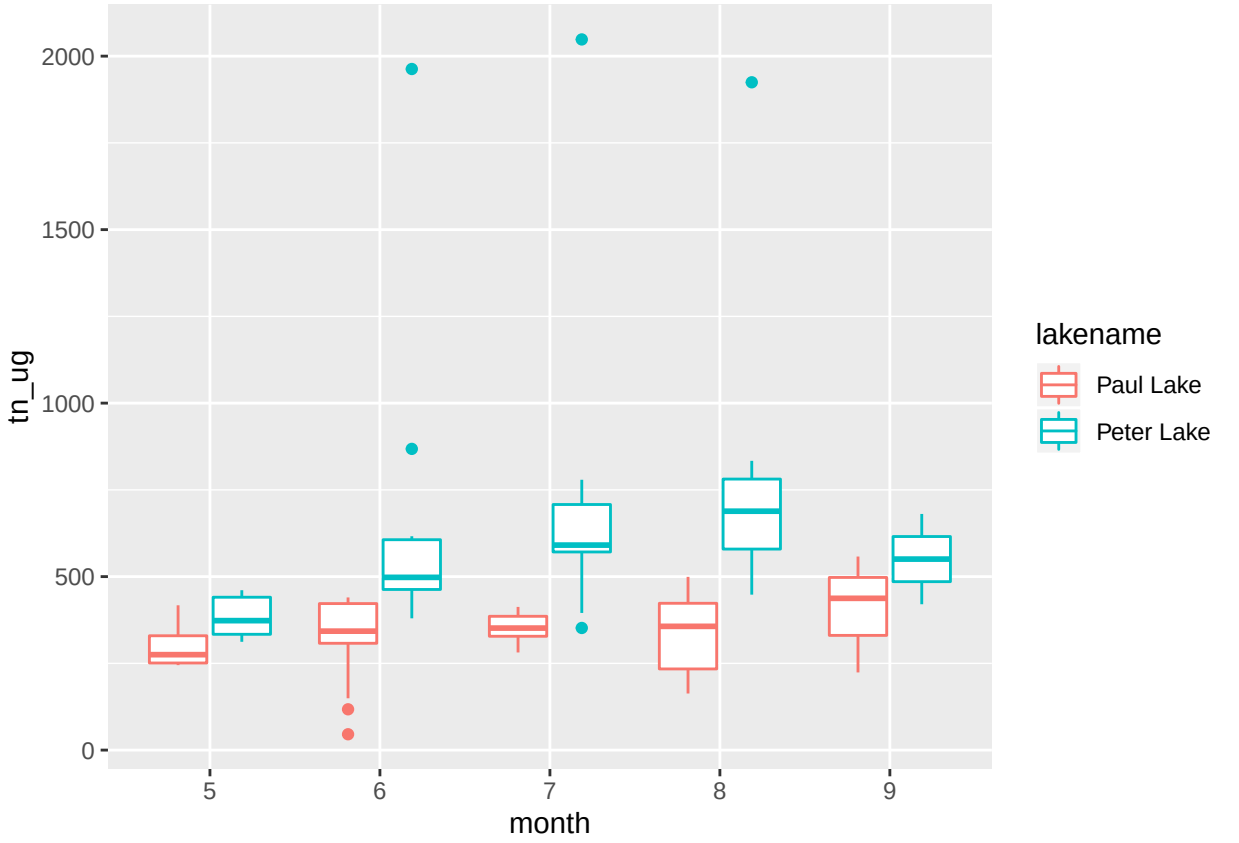
7. Create two plots, with TN (plot 1) or TP (plot 2) as the response variable and month and lake as the predictor variables. Hint: you may use some of the code you used for your visualization assignment. Assign groupings with letters, as determined from your tests. Adjust your axes, aesthetics, and color 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.

```
#7
# plot 1 (TN)
TN <- ggplot(subset, aes(x = month, y = tn_ug, color = lakename)) +
  geom_boxplot()
```

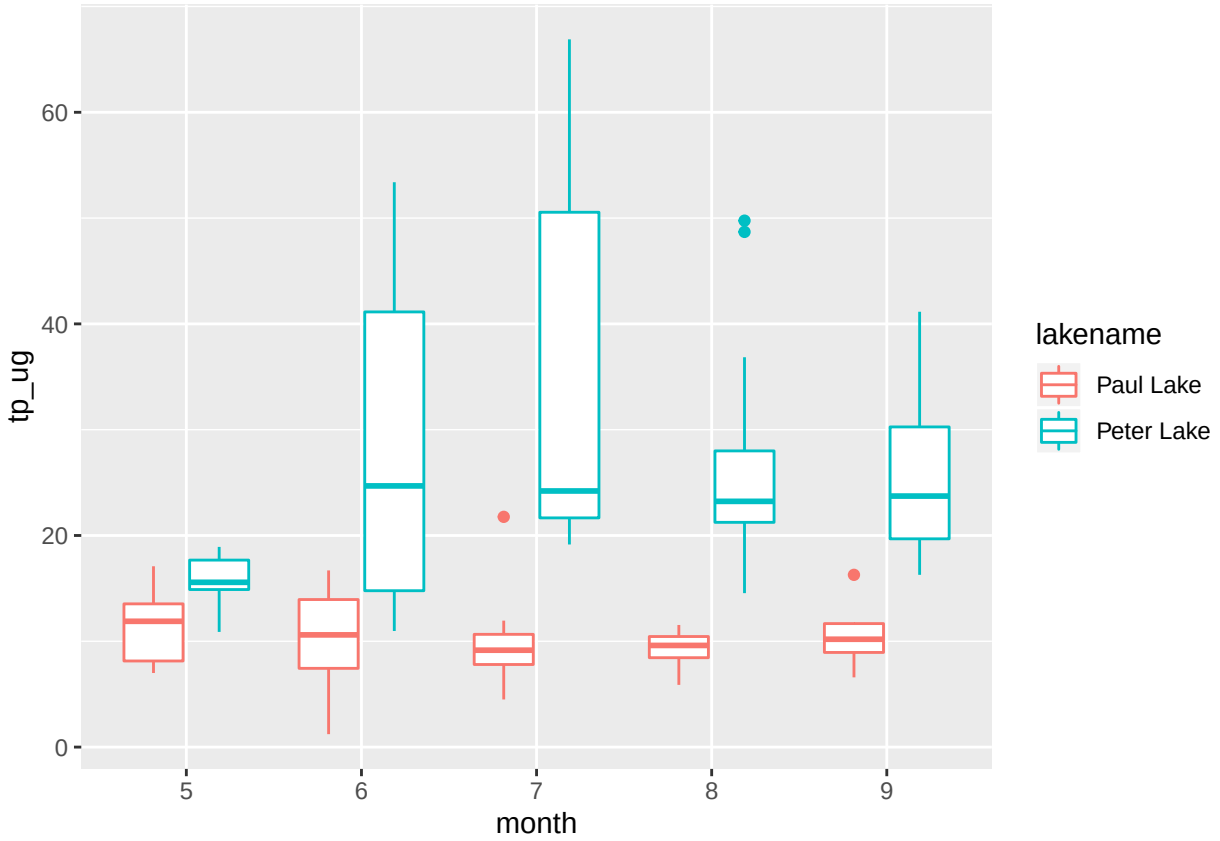
```
print(TN)
```

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




```
# plot 2 (TP)
TP <- ggplot(subset, aes(x = month, y = tp_ug, color = lakename)) +
  geom_boxplot()
print(TP)
```

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



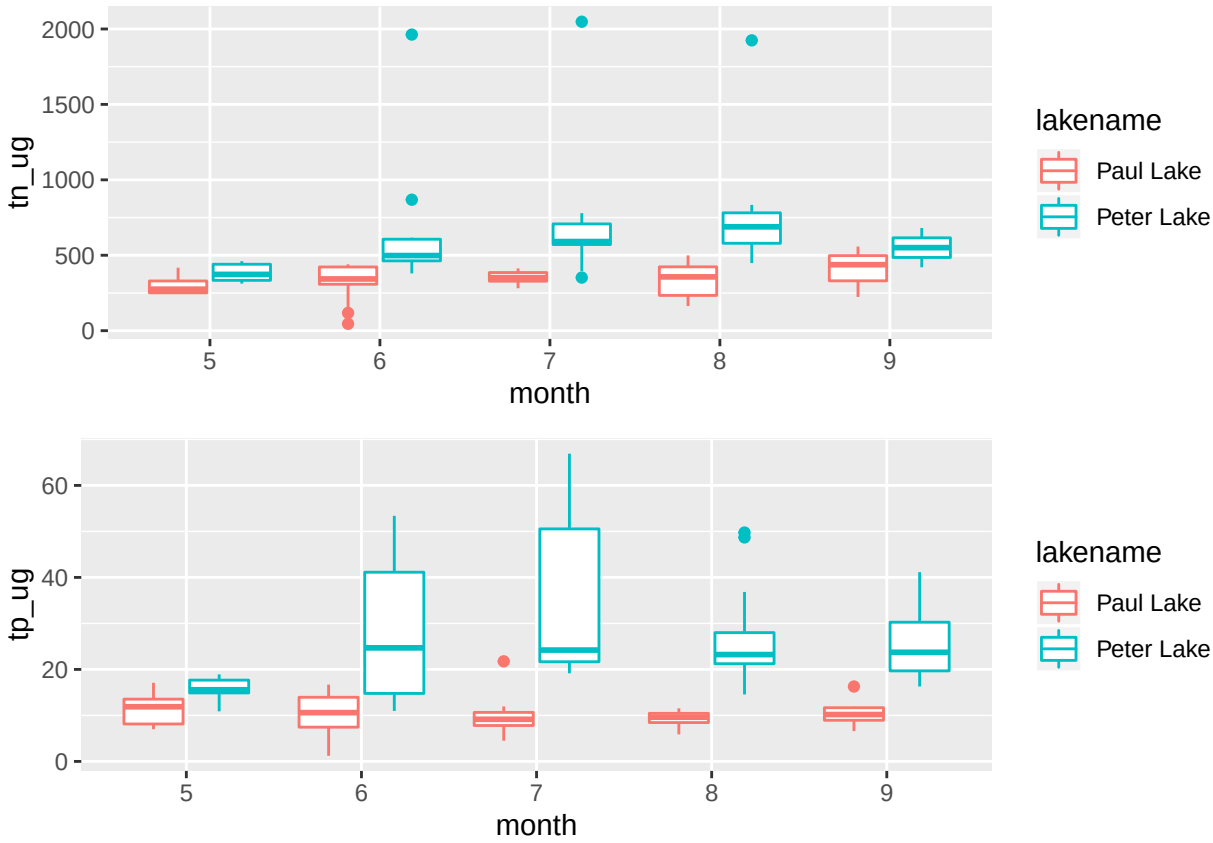

```
#8
# cowplot
legendA <- get_legend(TN + theme(legend.box.margin = margin(0, 0, 0, 12)))
```

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

```
combined <- plot_grid(TN, TP, legendA, ncol = 1, nrow = 2)
```

```
## Warning: Removed 23 rows containing non-finite values (stat_boxplot).
## Warning: Removed 1 rows containing non-finite values (stat_boxplot).
```

```
print(combined)
```


2000
1500
1000
500
0
tn_ug
5
6
7
8
9
month
lakename
Paul Lake
Peter Lake

60
40
20
0
tp_ug
5
6
7
8
9
month
lakename
Paul Lake
Peter Lake