What causes water pollution in a lake? : Analysis of water transparency and chemicals in Lake Ontario

Keisei_Aoki

12/5/2021

Introducion

The recent flood damaged the lower mainland. It reminds me that human activities are harming the ecosystem and environment, and one of the examples is the contamination and overfishing of the great lakes. When I was in elementary school, I learned that factors in polluting the sea are drainage chemicals, such as phosphorus and nitrogen. I wondered if the same factor also contaminates the great lakes as the sea. Therefore, I picked lake Ontario as a subject of the analysis. I analyzed the lake's water quality to estimate what factors pollute the lake and check whether my knowledge of sea pollution is consistent with the case of lake Ontario or not. So, I focused on the relationship between water quality and phosphorus and nitrogen in the analysis.

Data description

To analyze the water quality of Lake Ontario, I obtained data from the Government of Canada. The dataset is quite huge because there are 490 thousand rows and 21 columns. It captures a lot of information such as date of sampling, type of survey, the value of chemicals, names of chemicals etc. Also, over 400 different chemicals values were collected in the data and I picked water temperature, transparency, nitrogen, phosphorus, and oxygen as key variables out of numerous options. There is no explicit variable that measures level contamination, so I picked water transparency as a key variable because there is usually a high correlation between contamination of water transparency.

```
library(tidyverse)
library(broom)
data3 = read_csv("LAKE_ONTARIO_Water_Quality_2000-present.csv")
```

```
temp = data3 %>%
  filter(FULL_NAME ==  "TEMPERATURE (OF WATER)" ) %>%
  group_by(STN_DATE) %>%
  summarize(temp = mean(VALUE))
ph = data3 %>%
  filter(FULL_NAME == "PH" ) %>%
  group_by(STN_DATE) %>%
```

```
  summarize(ph = mean(VALUE))

clear = data3 %>%
  filter(FULL_NAME == "TRANSPARENCY" ) %>%
  group_by(STN_DATE) %>%
  summarize(transparency = mean(VALUE))
nh3  = data3 %>%
  filter(FULL_NAME == "AMMONIA NITROGEN,SOLUBLE") %>%
  group_by(STN_DATE) %>%
  summarize(nh3 = mean(VALUE))

other_nitrogen =  data3 %>%
  filter(FULL_NAME == "NITRATE+NITRITE NITROGEN,FILTERED") %>%
    group_by(STN_DATE) %>%
    summarize(other_n = mean(VALUE))

phosphorous  = data3 %>%
  filter(FULL_NAME == "PHOSPHOROUS,TOTAL" ) %>%
  group_by(STN_DATE) %>%
  summarize(phosphorous = mean(VALUE))

oxygen  =   data3 %>%
  filter(FULL_NAME == "OXYGEN,CONCENTRATION DISSOLVED") %>%
  group_by(STN_DATE) %>%
  summarize(oxygen = mean(VALUE))

data3 = data3 %>%
  filter(VALUE != -99.0000)

geo_data = data3 %>%
  select(STN_DATE,LATITUDE_DD,LONGITUDE_DD) %>%
  group_by(STN_DATE,LATITUDE_DD,LONGITUDE_DD) %>%
  summarize()

d1  = left_join(nh3,other_nitrogen,by="STN_DATE")
d1 = d1 %>%
    mutate(total_nitrogen = nh3 + other_n) %>%
    select(STN_DATE,total_nitrogen)
d1 = left_join(d1,ph,by="STN_DATE")
d2 = left_join(d1,temp,by="STN_DATE")
d3 = left_join(d2,oxygen,by="STN_DATE")
d4 = left_join(d3,phosphorous,by="STN_DATE")
d5 = left_join(d4,clear,by="STN_DATE")
d5 = left_join(d5,geo_data,by="STN_DATE")
d6 =  d5 %>%
          mutate(year = format(d5$STN_DATE,format="%Y"),
          month =format(d5$STN_DATE,format="%m")) %>%
      filter(year < 2018)
```

The main concern with the data set is tidiness because each row captures the different values of chemicals and other variables. The majority of data were collected in April, and the duration is 2000 to 2018, but not all variables were collected every year and on the same date. As a result, each chemical variable has a different number of total observations. However, the

number of observations is large enough, so it offsets the lacking consistency to some extent.

Summary statistics

Table

```
summary_stat = function(data) {
  c((length(data) - sum(is.na(data))),sd(data,na.rm = TRUE),
    min(data,na.rm = TRUE), as.vector(quantile(data,prob=c(.25),na.rm = TRUE)), mean(data,na.rm = TRUE)
max(data,na.rm = TRUE))
}

data_summary  = matrix(rep(1:7,5),ncol=7,byrow=TRUE)
data_summary[1,] = summary_stat(d6$transparency)
data_summary[2,] = summary_stat(d6$temp)
data_summary[3,] = summary_stat(d6$oxygen)
data_summary[4,] = summary_stat(d6$total_nitrogen)
data_summary[5,] = summary_stat(d6$phosphorous)

colnames(data_summary) <- c( "N","Std Dev", "Min","1st Qu","Mean","3rd Qu","Max")
rownames(data_summary) <- c("transparency","temperature","oxygen","Nitrogen",
                            "phosphorous")
names(dimnames(data_summary)) <- c("Key Variable", "Summary Statistics")

data_summary
```

```
##               Summary Statistics
## Key Variable     N      Std Dev      Min  1st Qu         Mean  3rd Qu      Max
##   transparency 825 13.683274862  2.00000 84.0000 83.905090909 90.0000  99.0000
##   temperature  848  1.679365325  0.50000  2.5000  3.586025943  4.5000  12.0000
##   oxygen       845  0.909979003 10.26500 12.8600 13.219231785 13.8132  15.7000
##   Nitrogen     851 23.378462502  0.09825  0.3533  1.301940426  0.4520 682.3222
##   phosphorous  579  0.006953687  0.00265  0.0058  0.008464072  0.0080   0.0724
```

```
d6 = d6%>%
  filter(total_nitrogen != 682.3222)
mean(d6$total_nitrogen)
```

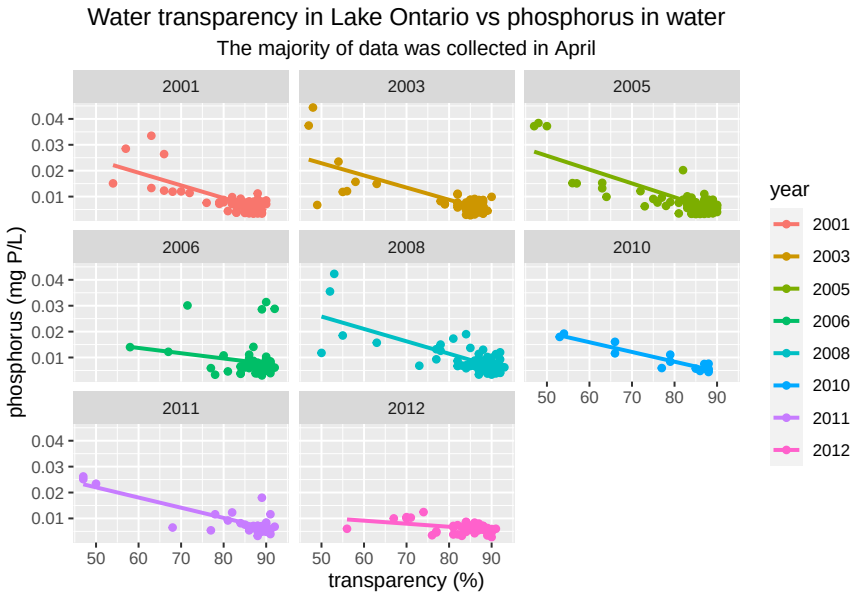
```
## [1] 0.5007401
```

The summary statistics of the key variables show a significant gap between the minimum and maximum values of nitrogen. The max value is 682.3222 mg/L. The extreme outlier pulls the mean value of nitrogen to the direction of outliers. Thus, I removed the outlier for this analysis. A mean value after removing the outlier is 0.5, which is less than half of the original mean. Other variables also have a big gap between minimum and maximum values, but it does not look like suffering from outliers much. Also, the statistic tells average water transparency is actually high.

Plot

```
d6 %>%
  filter ( year < 2013,transparency > 45) %>%
  ggplot(mappin=aes(x=transparency,y=phosphorous,group=year,color=year)) +
  geom_point() +
  geom_smooth(method = lm,se=FALSE)  +
  labs(title = "Water transparency in Lake Ontario vs phosphorus in water",
       subtitle = "The majority of data was collected in April",
       x="transparency (%)",
       y="phosphorus (mg P/L)") +
  theme(plot.title = element_text(hjust=0.5),
        plot.subtitle = element_text(hjust=0.5) ) +
       facet_wrap(~ year,nrow=3)
```

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



The scatter plot with a regression line shows that the relationship between water and phosphorus is negative. Despite each year having differences in correlation level, the plot shows the same trend all years. Which is the percentage of transparency tends to decrease as the amount of phosphorus in the water increases.

Linear regression model

```
regression = lm(transparency~total_nitrogen+phosphorous+temp+oxygen,data=d6)
```

I used water transparency as a dependent variable because I wanted to predict it in the regression model. Also, I used nitrogen, phosphorus, temperature and oxygen as independent

variables because I'd like to see the effect of these variables on water transparency to answer the questions. These questions are "What causes lake water pollution?" and "Do nitrogen and phosphorus cause such pollution?"

Table

```
reg_tidy = tidy(regression,conf.int=TRUE)
reg_tidy
```

```
## # A tibble: 5 x 7
##   term           estimate std.error statistic  p.value conf.low conf.high
##   <chr>             <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)      112.      12.2        9.17 9.34e-19    87.8    136.
## 2 total_nitrogen     5.54     1.22       4.55 6.74e- 6     3.14     7.93
## 3 phosphorous    -1652.      88.2      -18.7  7.89e-61 -1826.   -1479.
## 4 temp              -1.15     0.296     -3.88 1.15e- 4    -1.73    -0.569
## 5 oxygen            -1.05     0.845     -1.24 2.17e- 1    -2.71     0.615
```

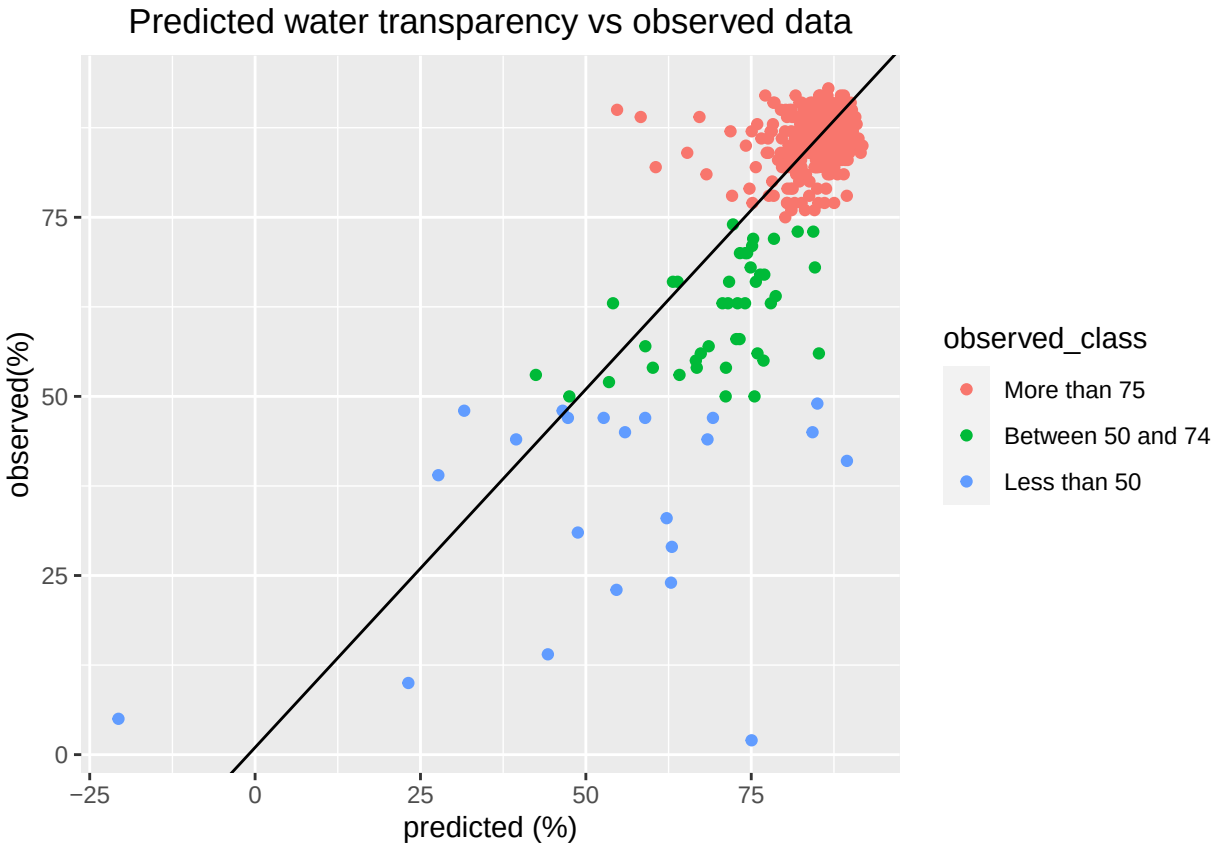
The regression result shows that all variables are statistically significant at a 5% confidence level except oxygen. The surprising result is that the regression predicts that nitrogen positively influences water transparency but on a small scale.

Plot

```
d6 = na.omit(d6) %>%
    mutate(predicion = fitted(regression),
    observed_class = if_else(transparency >= 75,"More than 75",""),
    observed_class = if_else((transparency < 75 & transparency
                              >=50),"Between 50 and 74",observed_class),
    observed_class = if_else(transparency <50,"Less than 50",observed_class),
    )

d6$observed_class = factor(d6$observed_class, levels = c("More than 75", "Between 50 and 74", "Less thar

ggplot(data=d6,mapping=aes(x=predicion,y=transparency,group=observed_class,color=observed_class)) +
  geom_point() +
  geom_abline(intercept = 1) +
  labs(title="Predicted water transparency vs observed data", x = "predicted (%)",
  y ="observed(%)") +
  theme(plot.title = element_text(hjust=0.5))
```



The plot shows the predicted value on the x-axis and the observed value on the y-axis. To easily see the accuracy of the prediction, I add the line with 0 intercepts and one slope. The plot tells that the regression model fits well when observed values are above 75% because most predictions are close to the observed values. However, the model performance worsens when the observed values get smaller and smaller. For example, a predicted value of 1 observation with values less than 10% is 75%. Also, another prediction of low observed value is -25% which is an unrealistic value.

Additional visualization/analysis

```
library(sf)
localarea_boundary <- st_read("Aquatic_resource_area_polygon_segment_.shp")
```

```
## Reading layer 'Aquatic_resource_area_polygon_segment_' from data source
##   '/Users/sasha/final-334/Aquatic_resource_area_polygon_segment_.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 131419 features and 0 fields
## Geometry type: POLYGON
## Dimension:     XY
## Bounding box:  xmin: -95.21303 ymin: 41.99537 xmax: -74.32011 ymax: 56.8567
## CRS:           NA
```

```
discription_boundary = read_csv("Aquatic_resource_area_polygon_segment_.csv")
```

Map

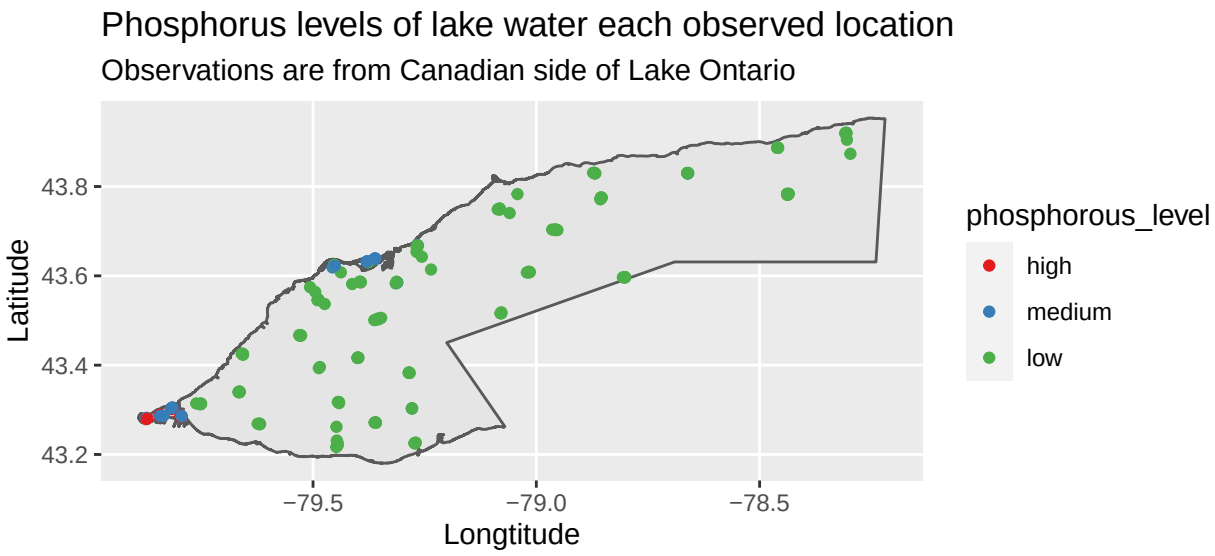
```
d8 = d6 %>%
  filter(LONGITUDE_DD < -79.2 & LATITUDE_DD > 43.2 |
           ((LONGITUDE_DD < -78.2 & LONGITUDE_DD > -79.2) &
              (LATITUDE_DD < 44   &  LATITUDE_DD > 43.5)))

d8 = na.omit(d8) %>%
  mutate(
     phosphorous_level =  if_else(phosphorous < 0.010,"low",""),
     phosphorous_level =  if_else((phosphorous >= 0.010
                          & phosphorous <= 0.035),"medium",phosphorous_level),
     phosphorous_level =  if_else(phosphorous > 0.035,"high",phosphorous_level))

d8$phosphorous_level =  factor(d8$phosphorous_level, levels = c("high", "medium", "low"))

zzz = localarea_boundary[128433,]

ggplot() + geom_sf(data=zzz) +
         geom_point(data=d8,mapping=aes(y=LATITUDE_DD,x=LONGITUDE_DD,group=phosphorous_level,color=phosp
      labs(title = "Phosphorus levels of lake water each observed location", subtitle="Observations are
```



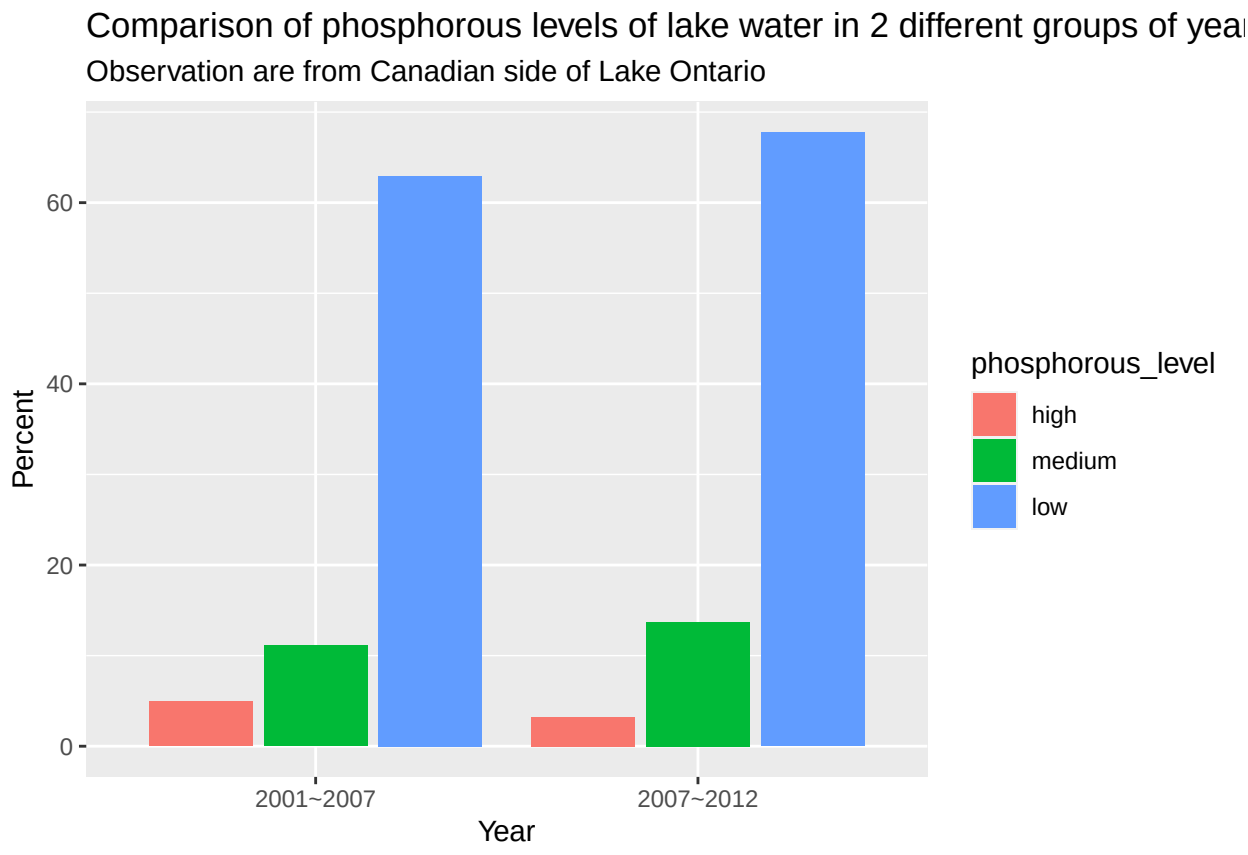
This is a plot of the map of the Canadian side of Lake Ontario near Toronto. Each point tells the water sampling location. To label phosphorus levels, I followed the general guide of the "trophic status" of the lake from the below link in the references part.

The map reveals that water near cities tend to have a higher phosphorous level than offshore because the location near Hamilton observed either high or medium level of phosphorous in the water. Also, some areas near Toronto show a medium level of the chemical. On the other hand, the water far from cities observed mostly a low phosphorous level.

Plot

```
d9 = d8 %>%
  mutate(
      year_group =  if_else(year <= 2007,"2001~2007",""),
       year_group =  if_else(year >= 2007,"2007~2012",year_group))
d90 = d9 %>%
  group_by(phosphorous_level,year_group) %>%
  summarise(N = n()) %>%
  mutate(freq = if_else(year_group == "2001~2007",N/(8+128+26),0),
         freq = if_else(year_group == "2007~2012",N/(4+95+25),freq),
         percent = round((freq*100), 2))

d90 %>%
    ggplot(mapping=aes(x=year_group,y=percent,group=phosphorous_level,fill=phosphorous_level)) + geom_c
```



The plot shows there is not much time difference in the proportion of each observed phosphorous level between 2 different periods. Therefore, it is hard to conclude whether the water quality of Lake Ontario is improving or not from the plot.

Conclusion

To conclude, the analysis shows the partially consistent result as what I learned in elementary school, which is phosphorous has a negative correlation with water quality, and the effect is relativity large. Also, there was evidence to support human activities are causing pollution because the phosphorus levels near cities are higher than offshore. On the other hand, it was surprising to see the regression model tells nitrogen has a positive relation to water quality. This indicates the regression might miss something, such as the correlation between variables. Thus, a cautious analysis is necessary to capture a real relationship.

References

Packaged used

- tidyverse
- broom
- sf

data names

- LAKE_ONTARIO_Water_Quality_2000-present.csv (https://open.canada.ca/data/en/dataset/cfdafa0c-a644-47cc-ad54-460304facf2e/resource/e08a76ad-59de-4006-8c4a-b4b64f48d3fc)
- Aquatic_resource_area_polygon_segment_.shx (https://geohub.lio.gov.on.ca/datasets/aquatic-resource-area-polygon-segment-/explore?location=49.291899%2C-84.834657%2C5.88)
- Aquatic_resource_area_polygon_segment_.csv (https://geohub.lio.gov.on.ca/datasets/aquatic-resource-area-polygon-segment-/explore?location=49.291899%2C-84.834657%2C5.88)

webpages/articles used

- https://socviz.co/workgeoms.html#workgeoms
- obtained general guide of "trophic status" of the lake from (https://www.knowyourh2o.com/outdoor-4/phosphates-in-the-environment)
- https://www.dummies.com/programming/r/how-to-remove-rows-with-missing-data-in-r/
- http://applied-r.com/rcolorbrewer-palettes/
- https://www.cyclismo.org/tutorial/R/tables.html
- https://bookdown.org/kdonovan125/ibis_data_analysis_r4/working-with-tables-in-r.html