Exploratory Data Analysis of City of Austin Water Quality

Contents

1) Introduction

Our group chose the Water Quality Sampling Data. After sifting through the variety of datasets offered to us by the city of Austin, we were drawn the most towards this dataset. We were interested in this data because of its relationship to the Austin environment. By analyzing this data, we can make informed decisions as civilians regarding our water usage and what we need to improve on in terms of city water quality. By using data analysis techniques and this dataset, we can become more educated about the source and quality of our drinking water and recreational bodies of water and track how water quality changes as the city continues to develop and grow over time.

Our main variables of interest in this dataset include DATE, WATERSHED, SITE_TYPE, PARAM_TYPE, PARAMETER, RESULT, and UNIT. Some of the question our data analysis will answer include:

1. Looking at the lakes, streams and springs in the greater Austin area, is there a difference between the results of water quality tests (Dissolved Carbon, Bacteria/Pathogen, pH, and Dissolved Oxygen) across these 3 different site types?
2. Across the three watershed Lake Austin, Lake Travis, and Lady Bird Lake, has the water quality in terms of dissolved oxygen changed over time overall in these 3 lakes?
3. Among the site types streams, lakes, and springs, is there a difference between what water quality tests are conducted and the number of tests and the site type?

2) Methods

```r
# For importing necessary library
library(readr)
library(tidyverse)
```

Loading the dataset so that we can begin exploring the data.

```r
# Loading the data set into a dataframe
water_quality_sampling_data <- read.csv('Water_Quality_Sampling_Data_20231020.csv')
```

Here we are just looking at and exploring the structure of the data to help us with cleaning later and selecting the variable we want to include in the final tidy dataset.

```r
# Looking at the structure of the data
str(water_quality_sampling_data)
```

```
## 'data.frame':    1440115 obs. of  23 variables:
##  $ DATA_REF_NO    : int  2308323 2308163 2308160 2308159 2308158 2308162 2308324 2308161 2308322 103
##  $ SAMPLE_REF_NO  : int  448739 448739 448739 448739 448739 448739 448739 448739 448739 88 ...
##  $ SAMPLE_DATE    : chr  "12/07/1947 12:00:00 AM" "12/07/1947 12:00:00 AM" "12/07/1947 12:00:00 AM"
##  $ TIME_NULL      : chr  "Y" "Y" "Y" "Y" ...
##  $ WATERSHED      : chr  "Barton Creek" "Barton Creek" "Barton Creek" "Barton Creek" ...
##  $ SAMPLE_SITE_NO : int  735 735 735 735 735 735 735 735 735 35 ...
##  $ SITE_NAME      : chr  "USGS Well 301526097463201 (Rabb Well)" "USGS Well 301526097463201 (Rabb Wel
##  $ LAT_DD_WGS84   : num  30.3 30.3 30.3 30.3 30.3 ...
##  $ LON_DD_WGS84   : num  -97.8 -97.8 -97.8 -97.8 -97.8 ...
##  $ SITE_TYPE      : chr  "Well" "Well" "Well" "Well" ...
##  $ PROJECT        : chr  "Groundwater" "Groundwater" "Groundwater" "Groundwater" ...
##  $ SAMPLE_ID      : chr  "" "" "" "" ...
##  $ DEPTH_IN_METERS: num  NA NA NA NA NA NA NA NA NA NA ...
##  $ MEDIUM         : chr  "Ground Water" "Ground Water" "Ground Water" "Ground Water" ...
##  $ PARAM_TYPE     : chr  "Nutrients" "Major Ions" "Major Ions" "Major Ions" ...
##  $ PARAMETER      : chr  "NITRATE AS N" "CHLORIDE" "SODIUM" "MAGNESIUM" ...
##  $ QUALIFIER      : chr  "" "" "" "" ...
##  $ RESULT         : num  1.5 14 10 20 55 20 241 244 220 0.24 ...
##  $ UNIT           : chr  "MG/L" "MG/L" "MG/L" "MG/L" ...
##  $ METHOD         : chr  "UNKNOWN" "UNKNOWN" "UNKNOWN" "UNKNOWN" ...
##  $ FILTER         : chr  "Total" "Total" "Total" "Total" ...
##  $ QC_TYPE        : chr  "" "" "" "" ...
##  $ QC_FLAG        : chr  "U" "U" "U" "U" ...
```

Cleaning the dataset

This dataset begins with 1440115 rows with 23 columns. We begin cleaning up the dataset for our anaylsis by selecting column variables we deem necessary for our data analysis. Since we are not interested in all the types of sites and test parameter types contained in the dataset, we filtered the dataset to only include the site types and test parameter types that we are interested in. After selecting all the relevant data for our analysis, we noticed that the SAMPLE_DATE column had values for both date and time so we separated them into their own columns and converted the dates into 'Date' data types for our time series analysis later. Once we performed all these steps, we found or data to be tidy since every observation had its own row and every variable had its own column. Our final tidy dataset contains 323437 rows and 15 variables.

```r
# Tidying/Wrangling/Manipulating dataset to get a clean/usable dataset

water_quality_sampling_data_tidy <- water_quality_sampling_data |>
    # selecting the column variables we are interested in our data analysis
    select(DATA_REF_NO, SAMPLE_REF_NO, SAMPLE_DATE, WATERSHED, SAMPLE_SITE_NO, SITE_NAME,
           LAT_DD_WGS84, LON_DD_WGS84, SITE_TYPE, DEPTH_IN_METERS, PARAM_TYPE, PARAMETER,
           RESULT, UNIT) |>
    # We are only interested in look at these Site types and test parameters
    filter(SITE_TYPE %in% c('Well', 'Spring', 'Stream', 'Lake', 'Cave Stream',
                            'Drinking Water Tap', 'Cave Pool','Cave Drip',
                            'Rainwater Catchment') &
           PARAM_TYPE %in% c('Nutrients', 'Major Ions', 'Solids/Conductivity',
                             'Alkalinity/Hardness/pH', 'Bacteria/Pathogens','Carbon',
                             'Metals', 'Oxygen', 'Hydrocarbons', 'Chlorinated',
                             'Insecticides', 'Fertilizers', 'Fungicides')) |>
    # Cleaning the SAMPLE_DATE column since it contains both the date and time
    separate(SAMPLE_DATE, into = c('DATE', 'TIME'), sep = ' ',
             convert = TRUE, extra = 'merge') |>
    # Converting the date information into a date object in R
    mutate(DATE = as.Date(DATE, format = '%m/%d/%Y'))
```

---

3) Results

Research Question 1: Looking at the lakes, streams and springs in the greater Austin area, is there a difference between the results of water quality tests (Dissolved Carbon, Bacteria/Pathogen, pH, and Dissolved Oxygen) across these 3 different site types?

Since most of our visualizations for this question are positively skewed, we have decided to use the median and IQR as our summary statistics of interest. Also we have removed outliers before calculating summary statistics and performing visualizations since outliers in the analysis of our researc question were either incorrectly input test results, made the visualization unreadable, or were values that were deemed outliers by outlier filtering.

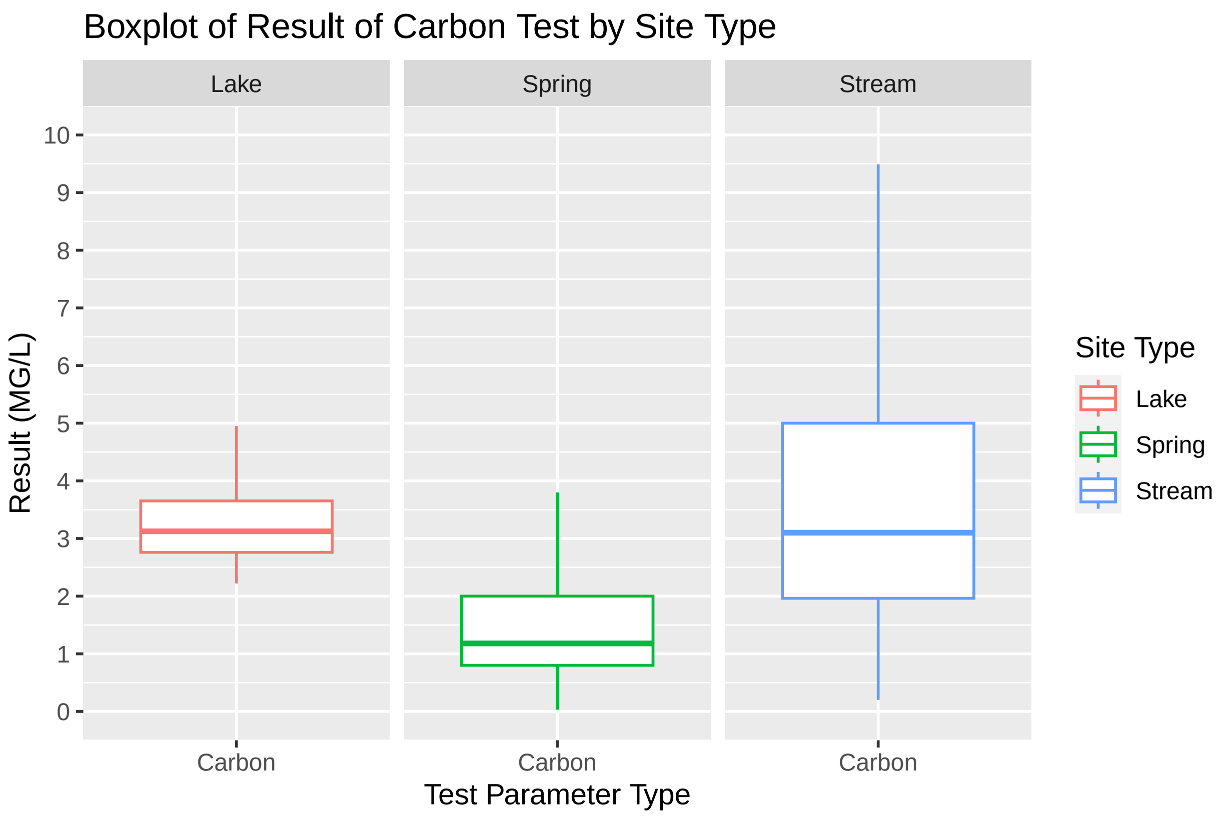
The site type stream has the highest median for carbon at 3.9 mg/L and an IQR of 6.02 mg/L. This is followed by the lake with a median of 3.210 mg/L and IQR of 1.225 mg/L and lastly spring with a median of 1.235 mg/L and IQR of 1.4325 mg/L. As you'll notice in future visualizations, the spring site type tends to have much lower averages and spreads than the other two site types.

```r
# Visualization 1 for research question 1

# summary statistics for the `RESULT` of the `Carbon` water quality test excluding outliers
water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Carbon' &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring') &
         PARAMETER == 'ORGANIC CARBON' &
         UNIT == 'MG/L') |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  summarize(
        Min = min(RESULT, na.rm = TRUE),
        Q1 = quantile(RESULT, 0.25, na.rm = TRUE),
        Median = median(RESULT, na.rm = TRUE),
        Mean = mean(RESULT, na.rm = TRUE),
        Q3 = quantile(RESULT, 0.75, na.rm = TRUE),
        Max = max(RESULT, na.rm = TRUE))
```

```
## # A tibble: 3 x 7
##   SITE_TYPE   Min    Q1 Median  Mean    Q3   Max
##   <chr>     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Lake       0.75  2.76   3.12  3.34  3.65  6.97
## 2 Spring     0.03  0.8    1.18  1.68  2    10.2
## 3 Stream     0.2   1.98   3.1   3.78  5.06 10.3
```

```r
# Generated a box plot for the result of the Carbon test across the 3 site types and removed outliers s

water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Carbon' &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring') &
         PARAMETER == 'ORGANIC CARBON' &
         UNIT == 'MG/L') |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  ggplot() +
        # included outlier.shape = NA to remove outliers from plot
        geom_boxplot(aes(x = PARAM_TYPE, y = RESULT, color = SITE_TYPE),
                     outlier.shape = NA) +
        scale_y_continuous(limits = c(0, 10), breaks = seq(0,10,1)) +
        labs(title = 'Boxplot of Result of Carbon Test by Site Type',
             x = 'Test Parameter Type',
             y = 'Result (MG/L)',
             color = 'Site Type',
             caption = 'visualization no.1') +
        facet_wrap(~SITE_TYPE,  nrow = 1)
```

```
## Warning: Removed 14 rows containing non-finite values (‘stat_boxplot()‘).
```



visualization no.1

The site type lake has the highest median for carbon at 3.125 mg/L and an IQR of 0.8925 mg/L. This is followed by the stream with a median of 3.1 mg/L and IQR of 3.08 mg/L and lastly spring with a median of 1.18 mg/L and IQR of 1.2 mg/L. As you'll notice in future visualizations, the spring site type tends to have much lower summary statistics and smaller spreads than the other two site types.

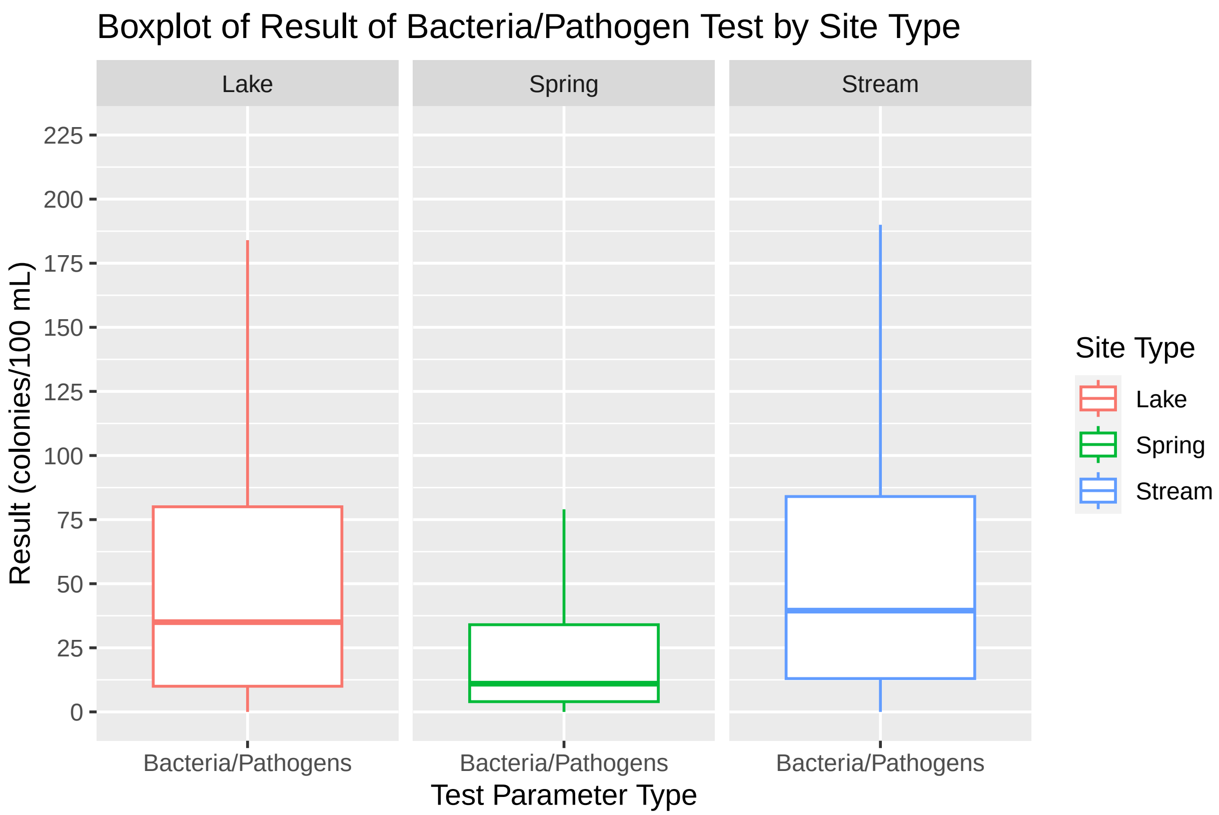
Based on the boxplot of visualization 1, the spring site type has a much lower carbon concentration in comparison to the lake and stream site type, as the spread doesn't overlap at all with lake or stream. The carbon concentration between the lake and stream are similar, but the carbon test results from the stream site type has a greater spread of values.

```
# Visualization 2 for research question 1

# summary statistics for the `RESULT` of the `Bacteria/Pathogen` water quality test excluding outliers
water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Bacteria/Pathogens' &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring') &
         UNIT == 'Colonies/100mL') |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  summarize(
        Min = min(RESULT, na.rm = TRUE),
        Q1 = quantile(RESULT, 0.25, na.rm = TRUE),
        Median = median(RESULT, na.rm = TRUE),
        Mean = mean(RESULT, na.rm = TRUE),
        Q3 = quantile(RESULT, 0.75, na.rm = TRUE),
        Max = max(RESULT, na.rm = TRUE))
```

```
## # A tibble: 3 x 7
##   SITE_TYPE   Min    Q1 Median  Mean    Q3   Max
##   <chr>     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Lake          0  10.8     42  74.9   100   410
## 2 Spring        0   4       12  40.2    40   410
## 3 Stream        0  15       45  80.1   108   413
```

```
# Generated a boxplot for the result of the Bacteria/Pathogen test across the 3 site types and removed

water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Bacteria/Pathogens' &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring') &
         UNIT == 'Colonies/100mL') |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  ggplot() +
        # included outlier.shape = NA to remove outliers from plot
        geom_boxplot(aes(x = PARAM_TYPE, y = RESULT, color = SITE_TYPE),
                     outlier.shape = NA) +
        scale_y_continuous(limits = c(0, 225), breaks = seq(0,225,25)) +
        labs(title = 'Boxplot of Result of Bacteria/Pathogen Test by Site Type',
             x = 'Test Parameter Type',
             y = 'Result (colonies/100 mL)',
             color = 'Site Type',
             caption = 'visualization no.2') +
        facet_wrap(~SITE_TYPE, nrow = 1)
```

```
## Warning: Removed 1456 rows containing non-finite values (‘stat_boxplot()‘).
```



visualization no.2

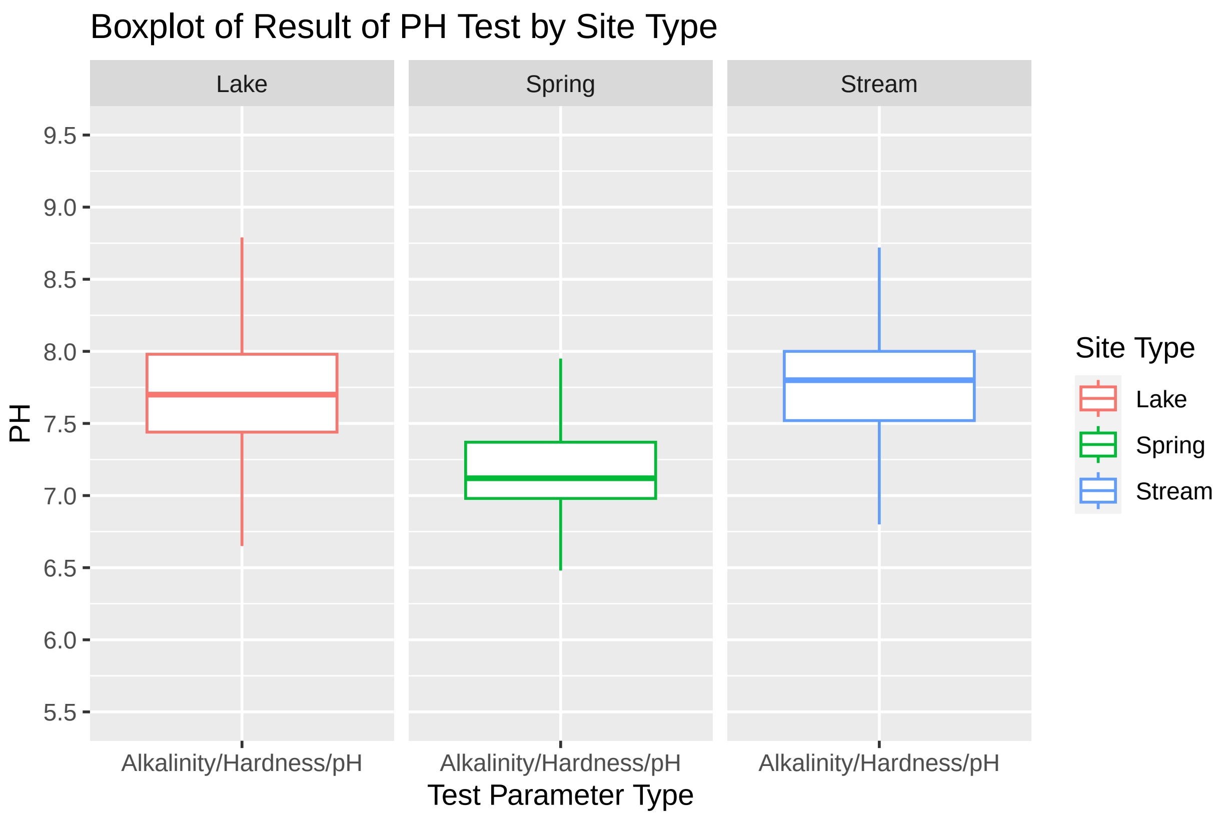
Looking at the summary statistics, the stream site type has the highest median again at 45 colonies/100 mL and a IQR of 93 colonies/100 mL. The lake site type has a median of 42 colonies/100 mL and a IQR of 89.25 colonies/100 mL. The lowest concentration is the spring site type at 12 colonies/100 mL with an IQR of 36 colonies/100 mL. As you can see, spring remains much lower than its counterparts.

Based on the boxplot of visualization 2, the spring site type again has the lowest median for the test results. There is also much overlap of spread between lake and stream site type, and the median and IQR are very similar. This maybe a result of the interconnections of lake and streams in Austin especially the streams that connect the major lakes together in Austin, which may result it similar test results.

```
# Visualization 3 for research question 1

# summary statistics for the `RESULT` of the `Alkalinity/Hardness/pH` specifically the pH water quality
water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Alkalinity/Hardness/pH' &
         PARAMETER %in% c('PH','24-HOUR AVG PH') &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring')) |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  summarize(
        Min = min(RESULT, na.rm = TRUE),
        Q1 = quantile(RESULT, 0.25, na.rm = TRUE),
        Median = median(RESULT, na.rm = TRUE),
        Mean = mean(RESULT, na.rm = TRUE),
        Q3 = quantile(RESULT, 0.75, na.rm = TRUE),
        Max = max(RESULT, na.rm = TRUE))
```

```
## # A tibble: 3 x 7
##   SITE_TYPE   Min    Q1 Median  Mean    Q3   Max
##   <chr>     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Lake       6.48  7.44   7.7   7.72  7.98  8.87
## 2 Spring     6.48  6.98   7.12  7.19  7.37  8.54
## 3 Stream     6.48  7.52   7.8   7.75  8     8.86
```

```
# Generated a boxplot for the result of the Alkalinity/Hardness/pH test across the 3 site types and rem

water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Alkalinity/Hardness/pH' &
         PARAMETER %in% c('PH','24-HOUR AVG PH') &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring')) |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  ggplot() +
        # included outlier.shape = NA to remove outliers from plot
        geom_boxplot(aes(x = PARAM_TYPE, y = RESULT, color = SITE_TYPE),
                     outlier.shape = NA) +
        scale_y_continuous(limits = c(5.5,9.5), breaks = seq(5.5,9.5,0.5)) +
        labs(title = 'Boxplot of Result of PH Test by Site Type',
             x = 'Test Parameter Type',
             y = 'PH',
             color = 'Site Type',
             caption = 'visualization no.3') +
        facet_wrap(~SITE_TYPE,  nrow = 1)
```



visualization no.3

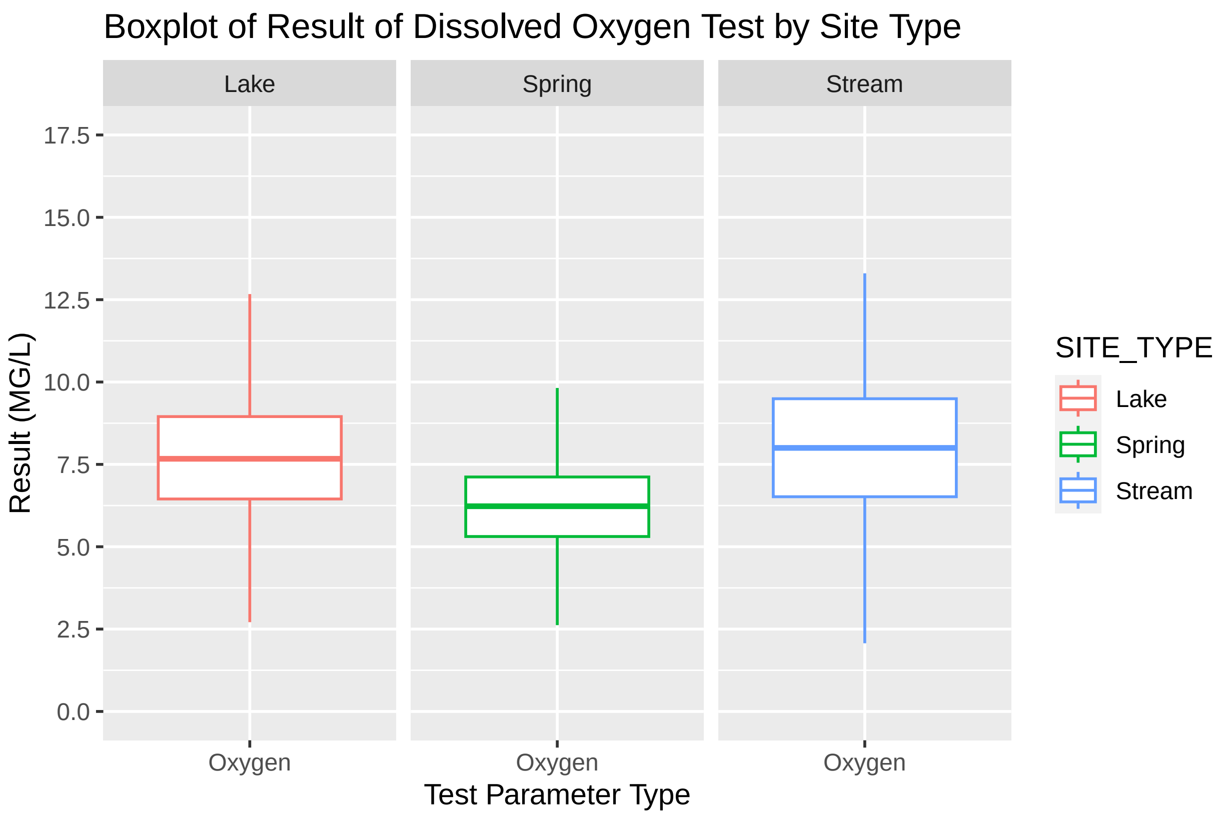
The pH concentration in all three site types is very similar and has a much smaller spread. The stream site type has the highest pH with a median of 7.80 and an IQR of 0.48. The lake site type has a median pH of 7.70 and an IQR of 0.54 and lastly, the spring site type has a median pH of 7.12 and an IQR of 0.39. The pH of water should be around 7, which is where all three site types hover around.

On the boxplot of visualization 3, the stream site type has the highest median pH out of all 3 site types while the spring site type has the lowest median pH compared to the other 2 sites, and it's below the first quartile of both lake and stream. It is worth noting that most of the test result data for the pH water quality tests lies within a neutral range pH which is to be expected for natural bodies of water, but it seems that lakes and streams are slightly more acidic than springs which makes sense since springs in Austin come from underground aquifers lined with hard limestone which can produce generally more basic water.

```r
# Visualization 4 for research question 1

# summary statistics for the `RESULT` of the `Oxygen` water quality test excluding outliers
water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Oxygen' &
         PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
         UNIT == 'MG/L' &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring')) |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>
  group_by(SITE_TYPE) |>
  summarize(
        Min = min(RESULT, na.rm = TRUE),
        Q1 = quantile(RESULT, 0.25, na.rm = TRUE),
        Median = median(RESULT, na.rm = TRUE),
        Mean = mean(RESULT, na.rm = TRUE),
        Q3 = quantile(RESULT, 0.75, na.rm = TRUE),
        Max = max(RESULT, na.rm = TRUE))
```

```
## # A tibble: 3 x 7
##   SITE_TYPE   Min    Q1 Median  Mean    Q3   Max
##   <chr>     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1 Lake       1.83  6.45   7.67  7.67  8.95  13.2
## 2 Spring     1.84  5.31   6.23  6.25  7.12  12.7
## 3 Stream     1.84  6.52   8     7.96  9.49  13.3
```

```
# Generated a boxplot for the result of the Oxygen test across the 3 site types and removed outliers so

water_quality_sampling_data_tidy |>
  filter(PARAM_TYPE == 'Oxygen' &
         PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
         UNIT == 'MG/L' &
         SITE_TYPE %in% c('Lake', 'Stream', 'Spring')) |>
  filter(RESULT >= quantile(RESULT, 0.25, na.rm = TRUE) - 1.5 * IQR(RESULT, na.rm = TRUE) &
         RESULT <= quantile(RESULT, 0.75, na.rm = TRUE) + 1.5 * IQR(RESULT, na.rm = TRUE)) |>   group_by
  ggplot() +
        # included outlier.shape = NA to remove outliers from plot
        geom_boxplot(aes(x = PARAM_TYPE, y = RESULT, color = SITE_TYPE),
                     outlier.shape = NA) +
        scale_y_continuous(limits = c(0, 17.5), breaks = seq(0,17.5,2.5)) +
        labs(title = 'Boxplot of Result of Dissolved Oxygen Test by Site Type',
             x = 'Test Parameter Type',
             y = 'Result (MG/L)',
             caption = 'visualization no.4') +
        facet_wrap(~SITE_TYPE,  nrow = 1)
```



visualization no.4

The stream site type continues to have the highest median with a value of 8.00 mg/L and IQR of 2.975 mg/L for dissolved oxygen. The lake site type has a median of 7.67 mg/L and IQR of 2.50 mg/L. The spring site type has a median of 6.23 mg/L and an IQR of 1.8075 mg/L. This follows the pattern we have observed with the other water quality test types we are analyzing. It worth highlighting that a healthy amount of dissolved oxygen is between 6.5 - 8 mg/L.

On the boxplot of visualization 4, the spring site type has a lower median and spread, and has a lower dissolved oxygen concentration in comparison to the lake and stream. Stream and lakes have very similar boxplot structure, indicating that the dissolved oxygen content across both site type may be similar.

Overall, based on the 4 water quality tests we analyzed it seems that generally the results of the spring water quality tests are different from the results of the lakes and stream. On the other hand, after looking at the boxplot across all tests we analyzed, it seems that lakes and streams post very similar water quality results and this may indicate some relationship between the 2 site types in terms of water quality.

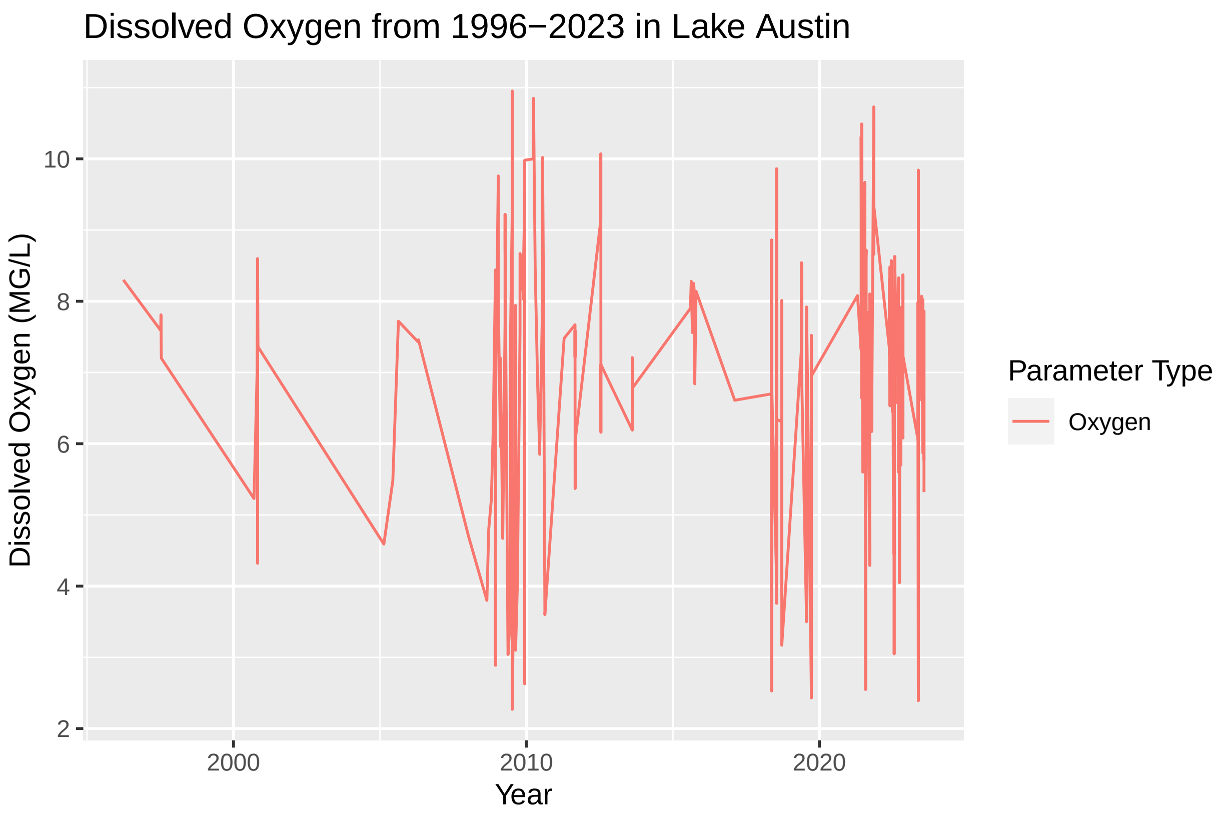
Research Question 2: Across the three watershed Lake Austin, Lake Travis, and Lady Bird Lake, has the water quality in terms of dissolved oxygen changed over time overall in these 3 lakes?

```
# Visualization 5 for research question 2 'Lake Austin'

water_quality_sampling_data_tidy |>
```

```r
    # filter for only Lake Austin with Oxygen quality test and with the correct units
    filter(WATERSHED == 'Lake Austin' &
           PARAM_TYPE %in% c('Oxygen') &
           PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
           UNIT == 'MG/L') |>
    select(RESULT) |>
    summary()
```

```
##      RESULT
##  Min.   :  0.000
##  1st Qu.:  5.695
##  Median :  7.270
##  Mean   :  8.229
##  3rd Qu.:  8.010
##  Max.   :841.000
```

```r
# We notice that there are outliers in the `RESULT` and proceed to remove them when visualizing

water_quality_sampling_data_tidy |>
    # filter for only Lake Austin with Oxygen quality test and with the correct units
    filter(WATERSHED == 'Lake Austin' &
           PARAM_TYPE %in% c('Oxygen') &
           PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
           UNIT == 'MG/L') |>
    # filtering again to remove the outliers
    filter(RESULT >= quantile(RESULT, 0.25, na.rm = T) - 1.5 * IQR(RESULT, na.rm = T) &
          RESULT <= quantile(RESULT, 0.75, na.rm = T) + 1.5 * IQR(RESULT, na.rm = T)) |>
    ggplot(aes(x = DATE, y = RESULT, color = PARAM_TYPE, group = PARAM_TYPE)) +
        geom_line() +
        labs(title = 'Dissolved Oxygen from 1996-2023 in Lake Austin',
             x = 'Year',
             y = 'Dissolved Oxygen (MG/L)',
             color = 'Parameter Type',
             caption = 'visualization no.5')
```



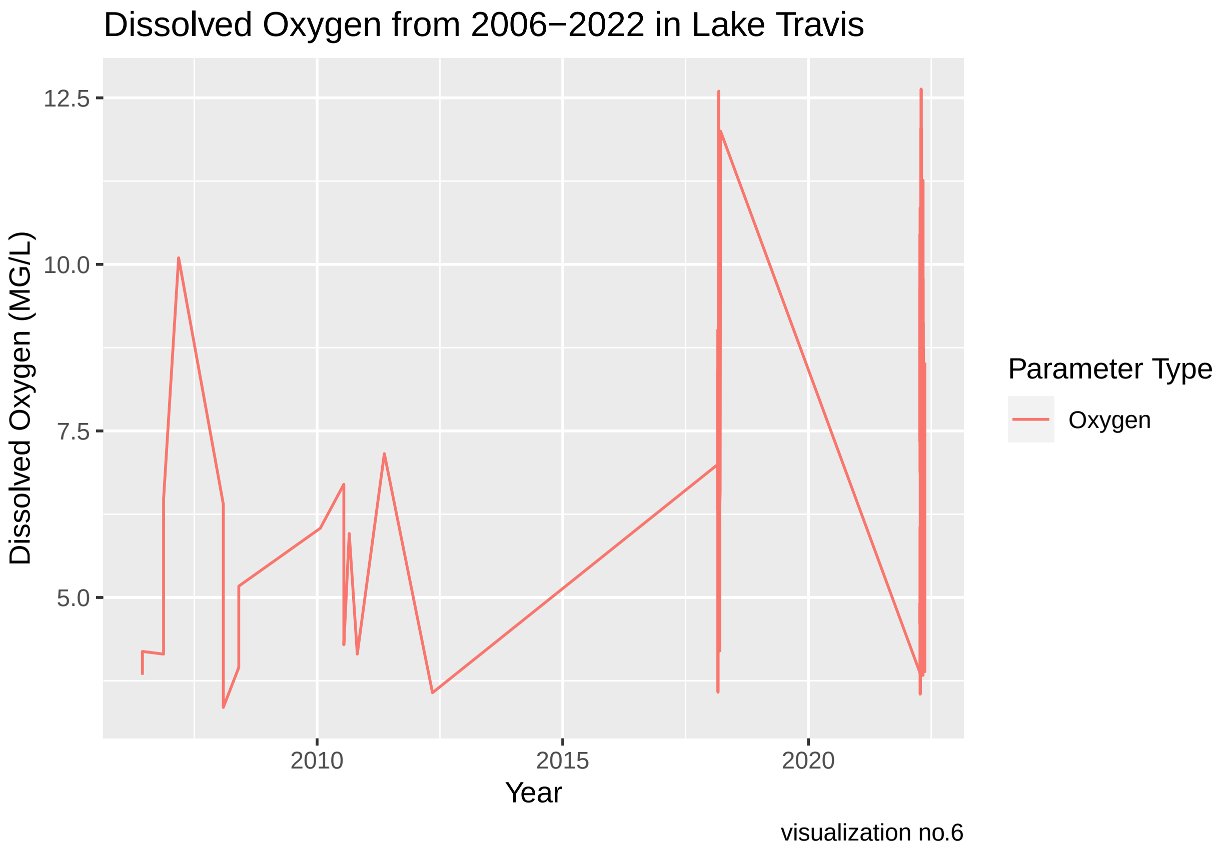
visualization no.5

Based on the summary statistics, the median over the years 1996 - 2023 of dissolved oxygen in Lake Austin is 7.27 mg/L. Looking at line graph of visualization 5, it seems that dissolved oxygen goes through cycles, fluctuating from very low to very high dissolved oxygen content likely due to seasonal changes, however, there doesn't seem to be an uptrend or downtrend of dissolved oxygen over the past 27 years of recorded data.

```
# Visualization 6 for research question 2 'Lake Travis'

water_quality_sampling_data_tidy |>
    # filter for only Lake Travis with Oxygen quality test and with the correct units
    filter(WATERSHED == 'Lake Travis' &
           PARAM_TYPE %in% c('Oxygen') &
           PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
           UNIT == 'MG/L') |>
    select(RESULT) |>
    summary()
```

```
##      RESULT
##  Min.   : 1.800
##  1st Qu.: 6.633
##  Median : 8.235
##  Mean   : 7.858
##  3rd Qu.: 9.115
##  Max.   :13.630
```

```r
# We notice that there are outliers in the `RESULT` and proceed to remove them when visualizing

water_quality_sampling_data_tidy |>
    # filter for only Lake Travis with Oxygen quality test and with the correct units
    filter(WATERSHED == 'Lake Travis' &
           PARAM_TYPE %in% c('Oxygen') &
           PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
           UNIT == 'MG/L') |>
    # filtering again to remove the outliers
    filter(RESULT >= quantile(RESULT, 0.25, na.rm = T) - 1.5 * IQR(RESULT, na.rm = T) &
          RESULT <= quantile(RESULT, 0.75, na.rm = T) + 1.5 * IQR(RESULT, na.rm = T)) |>
    ggplot(aes(x = DATE, y = RESULT, color = PARAM_TYPE, group = PARAM_TYPE)) +
        geom_line() +
        labs(title = 'Dissolved Oxygen from 2006-2022 in Lake Travis',
             x = 'Year',
             y = 'Dissolved Oxygen (MG/L)',
             color = 'Parameter Type',
             caption = 'visualization no.6')
```

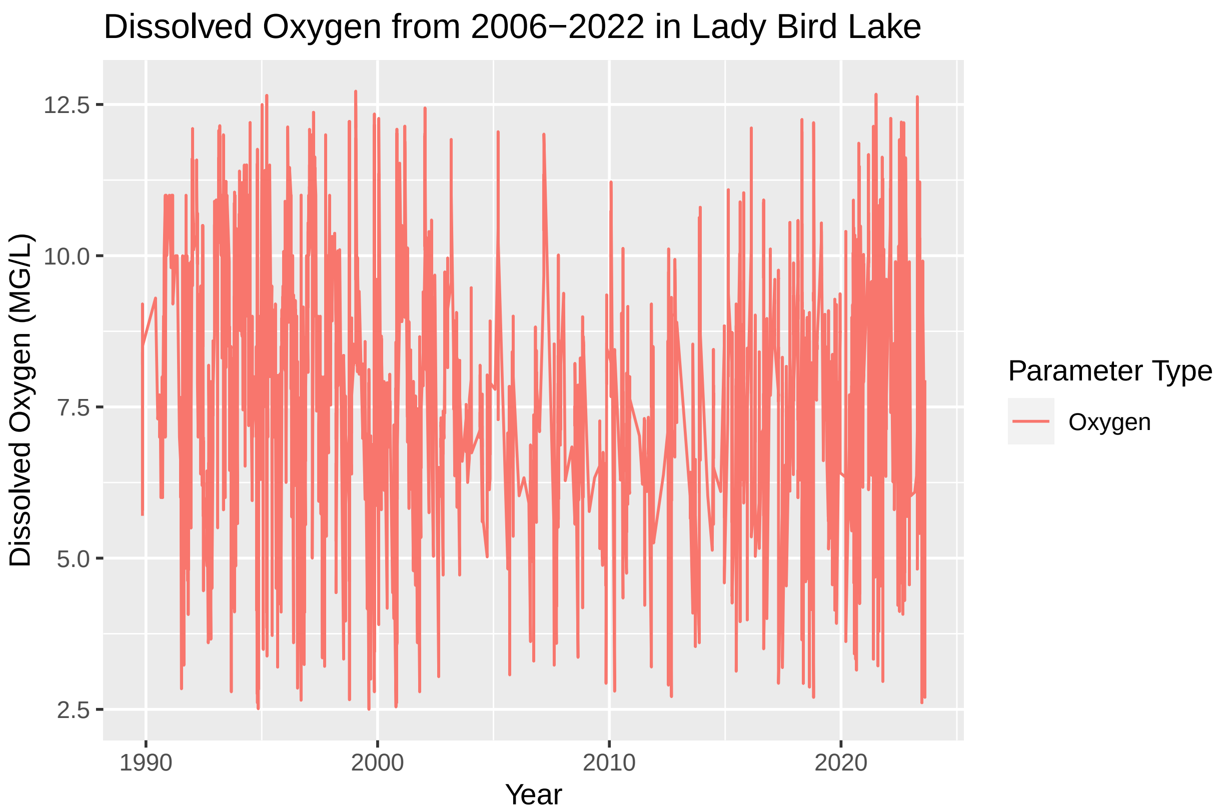


For Lake Travis, the median over the years 2006 - 2023 of dissolved oxygen is higher than Lake Austin, with a value of 8.235 mg/L. Based on the line graph of visualization 6, there is virtually nothing we can deduced as the data recorded on dissolved oxygen in Lake Travis is very sparse and sporadic and there is no clear trend we can see from the plot.

```r
# Visualization 7 for research question 2 'Lady Bird Lake'

water_quality_sampling_data_tidy |>
    # filter for only Lady Bird Lake with Oxygen quality test and with the correct units
    filter(WATERSHED == 'Lady Bird Lake' &
           PARAM_TYPE %in% c('Oxygen') &
           PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') &
           UNIT == 'MG/L') |>
    select(RESULT) |>
    summary()
```

```
##      RESULT
##  Min.   : 0.000
##  1st Qu.: 6.367
##  Median : 7.620
##  Mean   : 7.581
##  3rd Qu.: 8.950
##  Max.   :16.500
##  NA's   :3
```

```r
# We notice that there are outliers in the `RESULT` and proceed to remove them when visualizing

water_quality_sampling_data_tidy |>
    # filter for only Lady Bird Lake and Oxygen quality test and with the correct units
    filter(WATERSHED == 'Lady Bird Lake', PARAM_TYPE %in% c('Oxygen') &
            PARAMETER %in% c('24-HOUR AVG DISSOLVED OXYGEN','DISSOLVED OXYGEN') & UNIT == 'MG/L') |>
    # filtering again to remove the outliers
    filter(RESULT >= quantile(RESULT, 0.25, na.rm = T) - 1.5 * IQR(RESULT, na.rm = T) &
           RESULT <= quantile(RESULT, 0.75, na.rm = T) + 1.5 * IQR(RESULT, na.rm = T)) |>
    ggplot(aes(x = DATE, y = RESULT, color = PARAM_TYPE, group = PARAM_TYPE)) +
        geom_line() +
        labs(title = 'Dissolved Oxygen from 2006-2022 in Lady Bird Lake',
             x = 'Year',
             y = 'Dissolved Oxygen (MG/L)',
             color = 'Parameter Type',
             caption = 'visualization no.7')
```



visualization no.7

For Lady Bird Lake, the median over the years 2006 - 2022 of dissolved oxygen is very similar to Lake Austin's, with a value of 7.62 mg/L. Based on the line graph of visualization 7, there seems to be a pattern of cycling every year or so, which likely corresponds to seasonal changes and temperature, but again, there is no appearent uptrend or downtrend of dissolved oxygen over the past 16 years of recorded data.

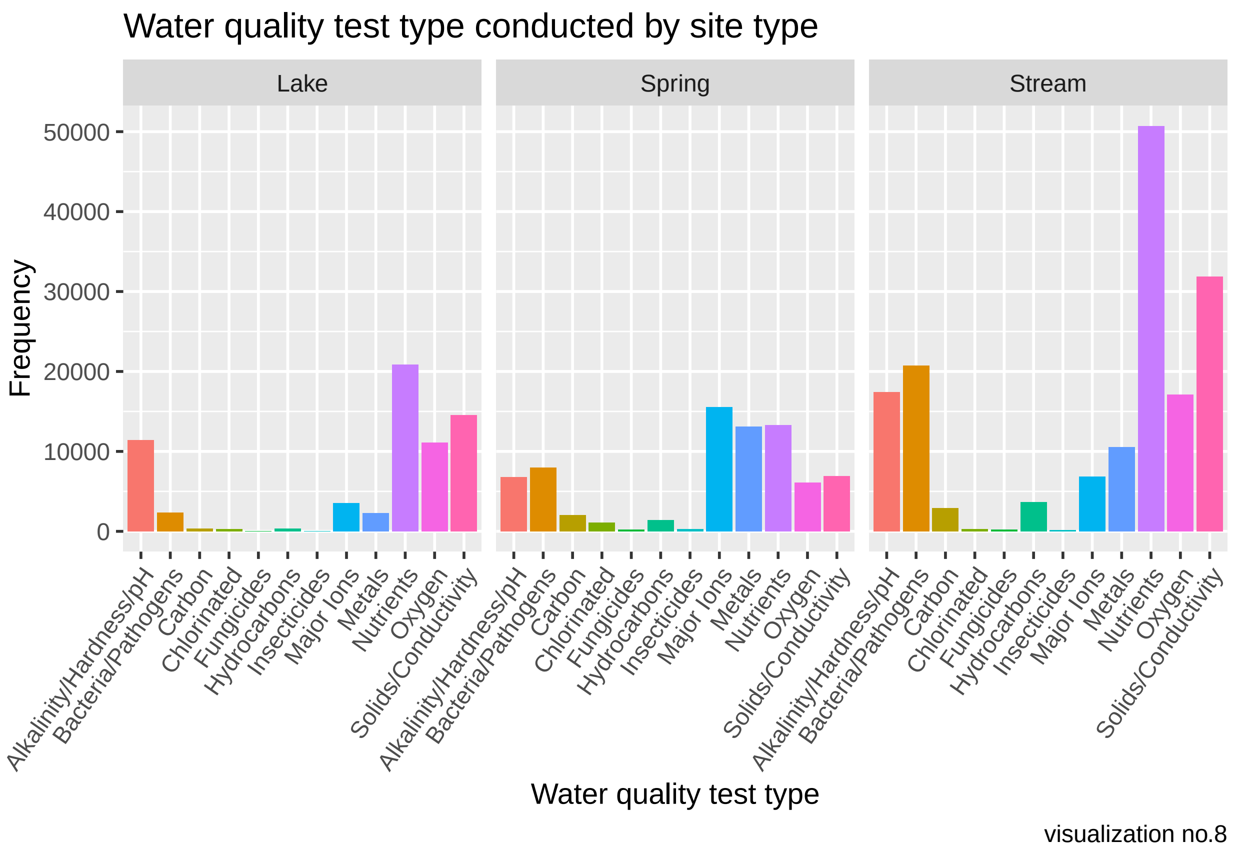
Overall, after analyzing the visualizations of dissolved oxygen content across the 3 lakes, it appears that over time the dissolved oxygen content doesn't change but rather follows a seasonal/cyclical pattern that decreases or increases the dissolved oxygen content within the lake through the year.

Research Question 3: Among the site types streams, lakes, and springs, is there a difference between what water quality tests are conducted and the number of tests and the site type?

```
# Visualization 8 for research question 3

water_quality_test_filtered <- water_quality_sampling_data_tidy |>
    filter(SITE_TYPE %in% c('Spring', 'Lake', 'Stream'))
table(water_quality_test_filtered$SITE_TYPE, water_quality_test_filtered$PARAM_TYPE)
```

```
##
##          Alkalinity/Hardness/pH Bacteria/Pathogens Carbon Chlorinated
##   Lake                    11404               2374    357         309
##   Spring                   6761               8001   2016        1110
##   Stream                  17425              20730   2902          281
##
##          Fungicides Hydrocarbons Insecticides Major Ions Metals Nutrients
##   Lake            8          350            4       3514   2262     20879
##   Spring        198         1423          301      15557  13075     13263
##   Stream        232         3690          149       6853  10532     50723
##
##          Oxygen Solids/Conductivity
##   Lake    11077               14559
##   Spring   6106                6926
##   Stream  17132               31897
```

```r
water_quality_sampling_data_tidy |>
    filter(SITE_TYPE %in% c('Spring', 'Lake', 'Stream')) |>
    ggplot() +
        geom_bar(aes(x = PARAM_TYPE, fill = PARAM_TYPE)) +
        labs(title = 'Water quality test type conducted by site type',
             x = 'Water quality test type',
             y = 'Frequency',
             caption = 'visualization no.8') +
        theme(legend.position='none', axis.text.x = element_text(angle = 55, hjust = 1)) +
        facet_wrap(~SITE_TYPE)
```



Looking at the bar chart of visualization 8, we can see that overall the stream site type had the most water quality tests performed, with nutrient tests being the type of tests most performed followed by oxygen and solids/conductivity tests. In comparison to the other 2 site types, lakes and springs, it seems that streams were tested more rigorously than the other 2 site types. Based on this bar graph, we can conclude that there was a major difference between the water quality test type conducted and the number of test for the streams site type but no so much for lakes and springs as the number of test and types of tests performed appear to be more similar across lakes and streams.

---

4) Discussion

From our data, we found that there was a relationship between the lake, stream and spring site type and the results of the water quality sample tests of dissolved carbon, bacteria, pH, and dissolved oxygen. Based on visualization 1, 2, 3 and 4, the spring site type seemed to consistently have a result of lower concentration of each water quality sample test. This is likely because springs are sourced from groundwater, whereas lakes and streams are sourced and function differently. Streams tended to have the highest concentration of each water quality sample test, which is interesting as they tend to start from elevated areas and are sourced largely from rainwater.

We also found that the water quality, in terms of dissolved oxygen, of the watersheds Lake Austin, Lake Travis, and Lady Bird Lake changed in volatile manners over the recorded years based on the visualizations 5-7. The concentration of dissolved oxygen has a daily/seasonal cycle, and is also affected by temperature which could explain the volatile nature of the dissolved oxygen concentrations for these watersheds across the years. The dissolved oxygen concentration can also tell us about the ability of aquatic life to prosper and for the body of water to be allowed for recreational use.

Finally, we found that there is a difference between the type of water quality tests conducted/the number of tests conduced and the site type looking at the plot of the spread of number of tests and type of test conducted by site type in visualization 8. The stream site type had the greatest number of observations conducted, indicating that streams in Austin are one of the most rigorously tested site types. Springs had the second-highest number of observations. It is extremely important that a high number of observations is gathered for springs and streams (and all bodies of water the city of Austin sources from) as it is often a source of drinking water or recreational use. Testing for contaminants is imperative to ensure the health of the greater Austin population. Springs are a source of freshwater, which only makes up 3% of the earth's water supply. It is important for us to maintain springs so that we do not reduce this already small percentage.

Our findings showcase the importance of being aware of water quality across the different bodies of water in Austin. Many locals enjoy the natural springs and pools that Austin has to offer. Fortunately, in our findings in research question one, we discovered that Austin maintains safe levels of the four main nutrients in lakes, springs, and streams. Dissolved oxygen levels tend to be quite volatile, but the median and IQR are healthy. Austin has collected a significant amount of samples for each site type, but we encourage them to increase this in future years, especially for springs, so that data is more consistent, trends are more defined and complete data analysis is more readily conducted. Our results convey for the most part that Austin has a clean and healthy supply of water for its users to drink from and enjoy on hot summer days.

---

5) Reflection, acknowledgements, and references

Some parts that were challenging about this project were the tidying process of the data and thinking about the best way to visualize the data. We had to determine the best variables to use for our data and remove the columns that weren't relevant to our research questions due to the sheer volume of observations included in this dataset. Initially, when we tried to visualize our data, there were too many observations to view the relationship between variables, especially with outliers. By filtering out certain variables and focusing on a specific number of site types and parameters, we could clearly see the relationships in the data.

Austine worked with cleaning the original dataset, and visualizing the data as well. GraceAnne and Seojin worked on analyzing the data and visualizations, and creating the presentation. We would like to thank the City of Austin for providing the water quality sampling data set for us to explore, and our professor Layla Guyot for giving us an opportunity to do this data analysis project.

Citations: https://www.austintexas.gov/department/water-quality-reports