EDA Report - Austin Water Quality Samples

Introduction

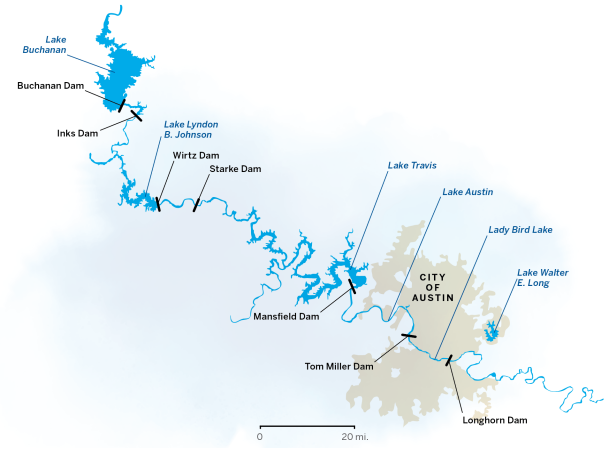


Figure 1: (Map of various water sources in the Austin Area. Image taken from the Jackson School of Geosciences at UT Austin.)

According to Austin Water Solutions (https://www.austinwatersolutions.net/understanding-austins-water-quality-issues/), Austin's water supply, just like many other cities', can be affected by many different issues, including contaminants from humans, bacteria, and leakage from aging water infrastructure. Austin's water quality affects everyone and every living thing in the city, and by maintaining a good water quality, we can ensure public health and prevent harmful contaminants from affecting residents, while also maintaining the environment in the area by protecting local ecosystems and aquatic life. To learn more about potential impurities and contaminants in Austin's water supply, I explored Austin's Water Quality Sampling dataset (https://data.austintexas.gov/Environment/Water-Quality-Sampling-Data/5tye-7ray/data_preview). Each row in the dataset represents one water sample taken. The main variables of interest include date, location, site type (stream, well, etc.), and different nutrients and pollutants the water was tested for, such as sodium, ammonia, and fecal bacteria.

My analysis of this dataset was based upon the following question: **How does the depth and location affect the filter result of an Austin water sample?**

I expected that different samples taken around the same location might share filter results, and that samples closer to the surface may have less contaminants, but overall, I didn't know what to expect walking into this dataset.

Methods

```
r #loading dataset and taking a look water_samples <-
filter(read_csv("Water_Quality_Sampling_Data_20250307.csv"), MEDIUM == "Surface Water")
head(water_samples)
```

```
## # A tibble: 6 x 24 ##   DATA_REF_NO SAMPLE_REF_NO PROJECT WATERSHED SITE_TYPE
SAMPLE_SITE_NO SITE_NAME ##         <dbl>         <dbl> <chr>   <chr>     <chr>
<dbl> <chr> ## 1     2429642        479356 Water ~ Barton C~ Stream               178
Barton C~ ## 2     2429638        479356 Water ~ Barton C~ Stream               178
Barton C~ ## 3     2418697        479356 Water ~ Barton C~ Stream               178
Barton C~ ## 4     2429641        479356 Water ~ Barton C~ Stream               178
Barton C~ ## 5     2429639        479356 Water ~ Barton C~ Stream               178
Barton C~ ## 6     2429862        479356 Water ~ Barton C~ Stream               178
Barton C~ ## # i 17 more variables: LAT_DD_WGS84 <dbl>, LON_DD_WGS84 <dbl>, ## #
SAMPLE_DATE <chr>, TIME_NULL <chr>, collecting_group <chr>, ## #   SAMPLE_ID <chr>,
MEDIUM <chr>, DEPTH_IN_METERS <dbl>, PARAM_TYPE <chr>, ## #   PARAMETER <chr>, QC_TYPE
<chr>, QUALIFIER <chr>, RESULT <dbl>, UNIT <chr>, ## #   QC_FLAG <chr>, METHOD <chr>,
FILTER <chr>
r nrow(water_samples)
## [1] 303730
r ncol(water_samples)
## [1] 24
```

This is the Austin Water Quality Samples dataset (from https://data.austintexas.gov/Environment/Water-Quality-Sampling-Data/5tye-7ray/data_preview), which records samples of water from various sources in Austin and their testing results. For my exploration, I will be focusing on surface water, rather than other mediums like groundwater, so I am filtering the dataset by water medium before saving. This dataset has 303730 rows and 24 columns.

```
r water_samples |> filter(is.na(DEPTH_IN_METERS) == TRUE)
## # A tibble: 206,950 x 24 ##    DATA_REF_NO SAMPLE_REF_NO PROJECT         WATERSHED
SITE_TYPE SAMPLE_SITE_NO ##           <dbl>         <dbl> <chr>           <chr>      <chr>
<dbl> ##  1     2429642        479356 Water Watch Dog Barton Cr~ Stream               178
##  2     2429638        479356 Water Watch Dog Barton Cr~ Stream               178 ##  3
2418697        479356 Water Watch Dog Barton Cr~ Stream               178 ##  4
2429641        479356 Water Watch Dog Barton Cr~ Stream               178 ##  5
2429639        479356 Water Watch Dog Barton Cr~ Stream               178 ##  6
2429862        479356 Water Watch Dog Barton Cr~ Stream               178 ##  7
2418696        479356 Water Watch Dog Barton Cr~ Stream               178 ##  8
2429863        479356 Water Watch Dog Barton Cr~ Stream               178 ##  9
2429640        479356 Water Watch Dog Barton Cr~ Stream               178 ## 10
2429489        479394 Water Watch Dog Barton Cr~ Stream                49 ## # i 206,940
more rows ## # i 18 more variables: SITE_NAME <chr>, LAT_DD_WGS84 <dbl>, LON_DD_WGS84
<dbl>, ## #   SAMPLE_DATE <chr>, TIME_NULL <chr>, collecting_group <chr>, ## #
SAMPLE_ID <chr>, MEDIUM <chr>, DEPTH_IN_METERS <dbl>, PARAM_TYPE <chr>, ## #   PARAMETER
<chr>, QC_TYPE <chr>, QUALIFIER <chr>, RESULT <dbl>, UNIT <chr>, ## #   QC_FLAG <chr>,
METHOD <chr>, FILTER <chr>
r water_samples |> filter(is.na(SAMPLE_ID) == TRUE)
```

```
## # A tibble: 145,978 x 24 ##    DATA_REF_NO SAMPLE_REF_NO PROJECT         WATERSHED
SITE_TYPE SAMPLE_SITE_NO ##          <dbl>         <dbl> <chr>           <chr>      <chr>
<dbl> ##  1     2429642        479356 Water Watch Dog Barton Cr~ Stream               178
##  2     2429638        479356 Water Watch Dog Barton Cr~ Stream               178 ##  3
2418697        479356 Water Watch Dog Barton Cr~ Stream               178 ##  4
2429641        479356 Water Watch Dog Barton Cr~ Stream               178 ##  5
2429639        479356 Water Watch Dog Barton Cr~ Stream               178 ##  6
2429862        479356 Water Watch Dog Barton Cr~ Stream               178 ##  7
2418696        479356 Water Watch Dog Barton Cr~ Stream               178 ##  8
2429863        479356 Water Watch Dog Barton Cr~ Stream               178 ##  9
2429640        479356 Water Watch Dog Barton Cr~ Stream               178 ## 10
2429489        479394 Water Watch Dog Barton Cr~ Stream                49 ## # i 145,968
more rows ## # i 18 more variables: SITE_NAME <chr>, LAT_DD_WGS84 <dbl>, LON_DD_WGS84
<dbl>, ## #   SAMPLE_DATE <chr>, TIME_NULL <chr>, collecting_group <chr>, ## #
SAMPLE_ID <chr>, MEDIUM <chr>, DEPTH_IN_METERS <dbl>, PARAM_TYPE <chr>, ## #   PARAMETER
<chr>, QC_TYPE <chr>, QUALIFIER <chr>, RESULT <dbl>, UNIT <chr>, ## #   QC_FLAG <chr>,
METHOD <chr>, FILTER <chr>
r #filtering and cleaning dataset samples_clean <- water_samples |>
filter(is.na(DEPTH_IN_METERS) == FALSE) |> filter(is.na(SAMPLE_ID) == FALSE) |>
mutate(SAMPLE_DATE = mdy_hms(SAMPLE_DATE)) |> rename("data_id" = "DATA_REF_NO",
"sample_id" = "SAMPLE_REF_NO", "project" = "PROJECT", "watershed" = "WATERSHED",
"site_type" = "SITE_TYPE", "site_number" = "SAMPLE_SITE_NO", "site" = "SITE_NAME",
"latitude" = "LAT_DD_WGS84", "longitude" = "LON_DD_WGS84", "date" = "SAMPLE_DATE",
"time_null" = "TIME_NULL", "group" = "collecting_group", "medium" = "MEDIUM", "depth" =
"DEPTH_IN_METERS", "type" = "PARAM_TYPE", "parameter" = "PARAMETER", "qc_type" =
"QC_TYPE", "qualifier" = "QUALIFIER", "result" = "RESULT", "unit" = "UNIT", "flag" =
"QC_FLAG", "method" = "METHOD", "filter" = "FILTER") nrow(samples_clean)
## [1] 46920
r ncol(samples_clean)
## [1] 24
r samples_clean
## # A tibble: 46,920 x 24 ##    data_id sample_id project      watershed site_type
site_number site  latitude ##      <dbl>     <dbl> <chr>        <chr>     <chr>
<dbl> <chr>    <dbl> ##  1 1867306    358887 Barton Spri~ Barton C~ Stream           4974
Bart~     30.3 ##  2 1867132    358888 Barton Spri~ Barton C~ Stream           4974 Bart~
30.3 ##  3 1867131    358889 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  4 1867130    358890 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ##  5
1866389    358947 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ##  6
1867129    358891 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ##  7
1867128    358892 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ##  8
1867127    358893 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ##  9
1867126    358894 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ## 10
1867125    358895 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3 ## # i
46,910 more rows ## # i 16 more variables: longitude <dbl>, date <dttm>, time_null <chr>,
## #   group <chr>, SAMPLE_ID <chr>, medium <chr>, depth <dbl>, type <chr>, ## #
parameter <chr>, qc_type <chr>, qualifier <chr>, result <dbl>, unit <chr>, ## #   flag
<chr>, method <chr>, filter <chr>
```

There are 206,950 rows with no available depth and 145,978 rows with no available sample ID. I filtered out these values because I would not be able to use values without depth for my exploration and rows without sample ID could not confirm the data in those rows.

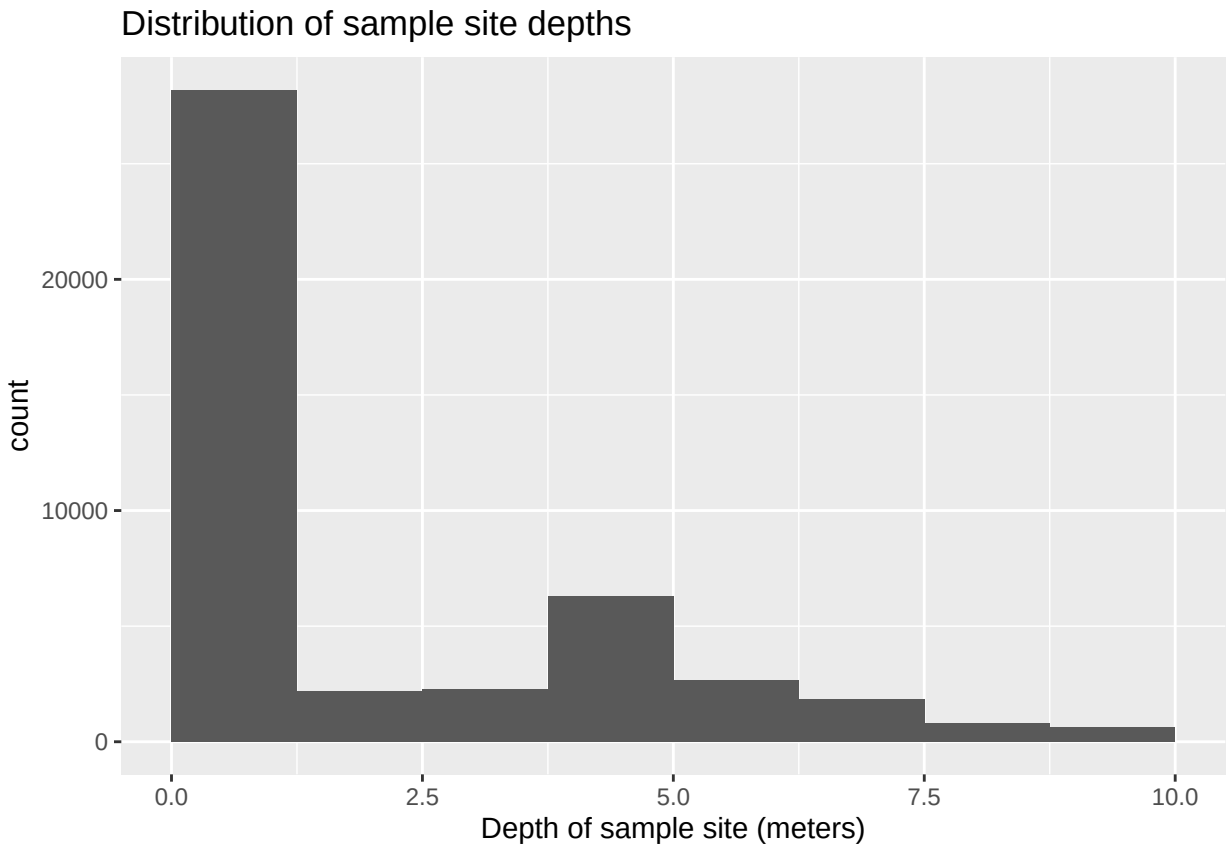
After wrangling, I ended up with a set of 46,920 rows and 24 columns to analyze. Because of the filtering efforts, the dataset is now much shorter and easier to process and work with. I also changed the sample date value from a character to a date datatype to make it easier to work with. This resulting dataset is tidy because every sample observation has its own row, each variable has its own column, and each data value has its own cell.

Analyzing Individual Variable: Depth

```
#creating visualization and statistics
summary(samples_clean$depth)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.100   0.300   0.300   2.368   4.000  18.000
```

```
ggplot(data = filter(samples_clean, depth < (1.5 * IQR(samples_clean$depth) + 4))) +
  geom_histogram(aes(x = depth), binwidth = 1.25, center = 0.625) +
  labs(title = "Distribution of sample site depths", x = "Depth of sample site (meters)")
```



The minimum depth is 0 meters and the maximum depth is 18 meters. The median depth is 0.3 meters and the mean depth is 2.368 meters. Judging from the visualization, there is a strong right skew in the data due to a large amount of outliers, even though most of the depth values are between 0 and 5 meters.

```r
#creating visualization and statistics
samples_clean |>
  group_by(site)|>
  summarize(largest_depth = max(depth)) |>
  slice_max(n = 10, largest_depth)
```

```
## # A tibble: 12 x 2
##    site                                              largest_depth
##    <chr>                                                     <dbl>
##  1 Lake Long @ Dam (LWL3)                                     18
##  2 Lake Austin @ Tom Miller Dam (AC)                          14.7
##  3 Lake Long mid-lake of Western Arm (LWL5)                   12.4
##  4 Lake Long mid-lake of Eastern Arm (LWL4)                   12.2
##  5 Lake Austin @ Emma Long Park (CC)                           9.4
##  6 Lady Bird Lake @ Basin [AC]                                 8
##  7 Lady Bird Lake Fish Survey Transect Midpoint IV             7.4
##  8 Lake Austin Fish Survey Transect Midpoint IV                7
##  9 Lady Bird Lake @ 1st St [CC]                                6.1
## 10 Lady Bird Lake @ Red Bud Isle (EC)                          6
## 11 Lake Austin Fish Survey Transect Midpoint I                 6
## 12 Lake Long @ Plant Intake (LWL1)                             6
```

Analyzing each site by the deepest point a sample was taken from (since samples can be taken at multiple points), we can see the top 10 deepest sample sites were all at Austin lakes.

Analyzing Individual Variable: Location (based on latitude and longitude)

```r
#creating visualization based on latitude and longitude
samples_clean |>
  group_by(site)
```

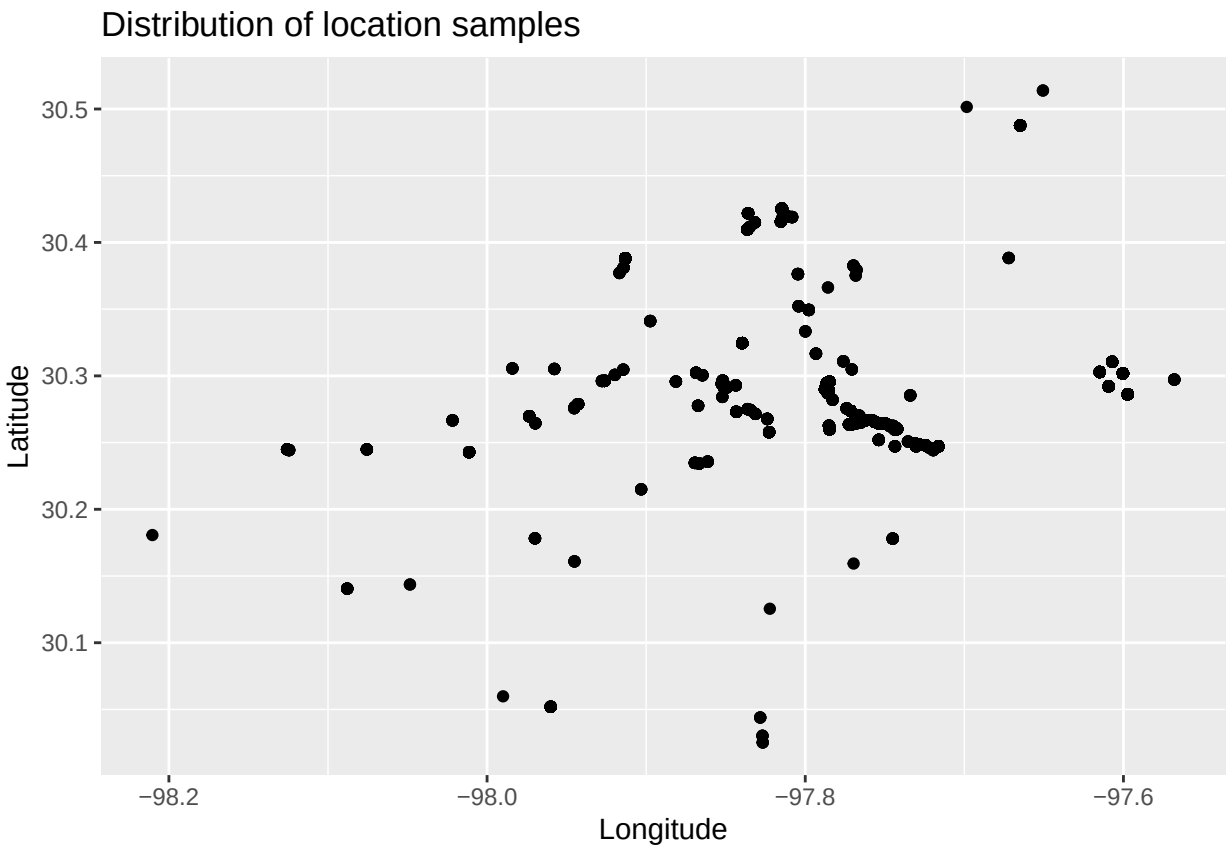
```
## # A tibble: 46,920 x 24
## # Groups:   site [135]
##    data_id sample_id project      watershed site_type site_number site  latitude
##      <dbl>     <dbl> <chr>        <chr>     <chr>           <dbl> <chr>    <dbl>
##  1 1867306    358887 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  2 1867132    358888 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  3 1867131    358889 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  4 1867130    358890 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  5 1866389    358947 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  6 1867129    358891 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  7 1867128    358892 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  8 1867127    358893 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
##  9 1867126    358894 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
## 10 1867125    358895 Barton Spri~ Barton C~ Stream           4974 Bart~     30.3
## # i 46,910 more rows
## # i 16 more variables: longitude <dbl>, date <dttm>, time_null <chr>,
## #   group <chr>, SAMPLE_ID <chr>, medium <chr>, depth <dbl>, type <chr>,
## #   parameter <chr>, qc_type <chr>, qualifier <chr>, result <dbl>, unit <chr>,
## #   flag <chr>, method <chr>, filter <chr>
```

```
samples_clean |>
  group_by(site)|>
  group_by(sample_id)|>
  summarize(count = n()) |>
  summary(count)
```

```
##    sample_id          count
##  Min.   :  1134   Min.   : 1.000
##  1st Qu.:358926   1st Qu.: 6.000
##  Median :556163   Median : 6.000
##  Mean   :462710   Mean   : 7.315
##  3rd Qu.:588468   3rd Qu.: 8.000
##  Max.   :597139   Max.   :24.000
```

```
samples_clean |>
  ggplot()+
  geom_point(aes(x = longitude, y = latitude)) +
  labs(title = "Distribution of location samples", x = "Longitude", y = "Latitude")
```

```
## Warning: Removed 26 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



In this dataset, there are 135 separate location sites that have been sampled. The minimum amount of separate times that a location site was sampled was 1 time, and the maximum

amount a site was sampled was 24 times. The median amount that a site was sampled was 6 times, while the mean amount a site was sampled was 7.315 times. From the scatterplot, we can see the distribution of location: the location sites are distributed fairly evenly across Austin, but there is some clustering around (-97.8, 30.25).

Analyzing Variable Sets: Filter Results vs. Location

```
samples_clean |>
  group_by(site) |>
  filter(parameter == "FECAL COLIFORM BACTERIA")|>
  summarize(fecal_bacteria_average = mean(result)) |>
  summary(fecal_bacteria_average)
```

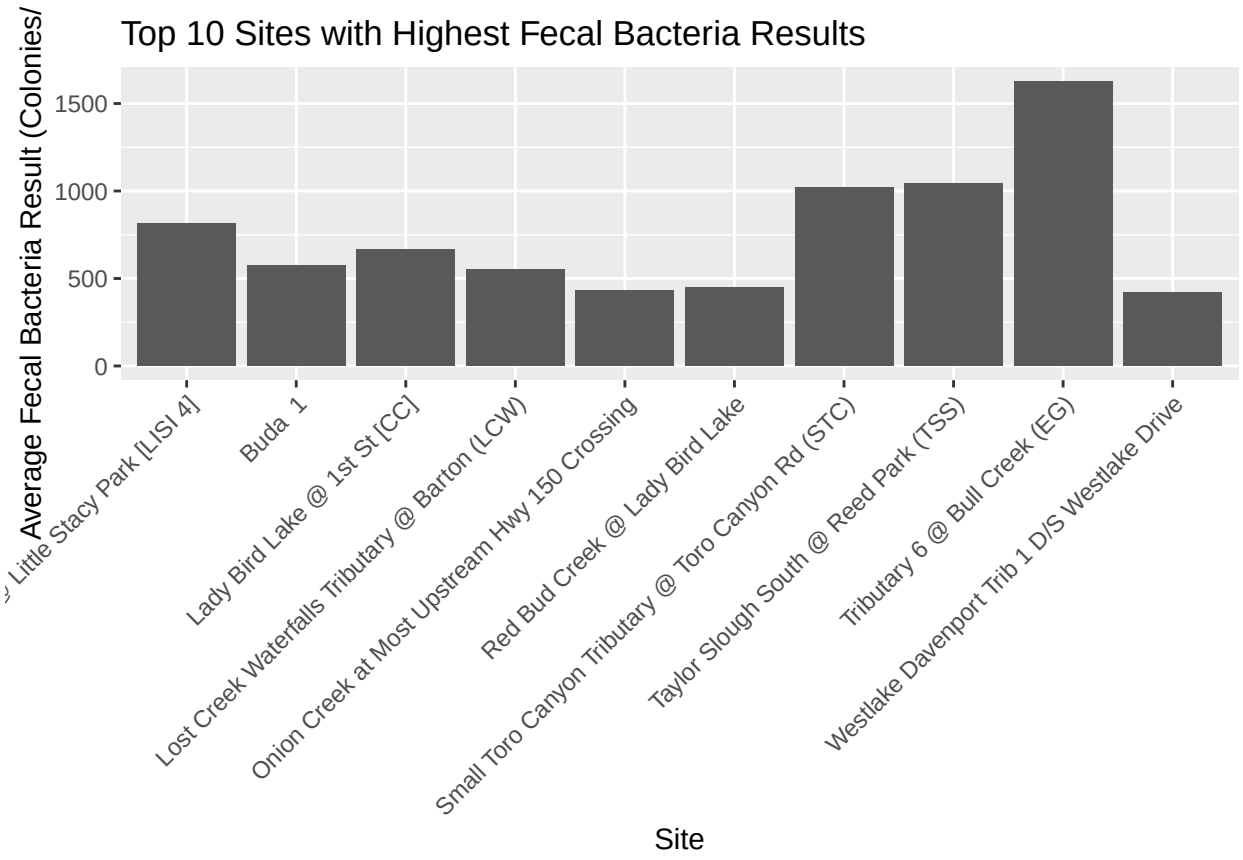
```
##      site           fecal_bacteria_average
##  Length:73          Min.   :   4.0
##  Class :character   1st Qu.:  43.0
##  Mode  :character   Median : 108.0
##                     Mean   : 204.5
##                     3rd Qu.: 234.5
##                     Max.   :1624.3
##                     NA's   :1
```

```
samples_clean |>
  group_by(site) |>
  filter(parameter == "FECAL COLIFORM BACTERIA")|>
  summarize(fecal_bacteria_average = mean(result)) |>
  slice_max(n = 10, fecal_bacteria_average)
```

```
## # A tibble: 10 x 2
##    site                                                fecal_bacteria_average
##    <chr>                                                                <dbl>
##  1 Tributary 6 @ Bull Creek (EG)                                        1624.
##  2 Taylor Slough South @ Reed Park (TSS)                                1044
##  3 Small Toro Canyon Tributary @ Toro Canyon Rd (STC)                   1022
##  4 Blunn Creek @ Little Stacy Park [LISI 4]                              816
##  5 Lady Bird Lake @ 1st St [CC]                                          665.
##  6 Buda  1                                                               576
##  7 Lost Creek Waterfalls Tributary @ Barton (LCW)                        550
##  8 Red Bud Creek @ Lady Bird Lake                                        450
##  9 Onion Creek at Most Upstream Hwy 150 Crossing                         430
## 10 Westlake Davenport Trib 1 D/S Westlake Drive                          420
```

```
samples_clean |>
  group_by(site) |>
  filter(parameter == "FECAL COLIFORM BACTERIA")|>
  summarize(fecal_bacteria_average = mean(result)) |>
  slice_max(n = 10, fecal_bacteria_average) |>
  ggplot()+
```

```
  geom_col(aes(x = site, y = fecal_bacteria_average)) +
  labs(x = "Site", y = "Average Fecal Bacteria Result (Colonies/100mL)", title = "Top 10 Sites with Hig
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Filtering to observe only the average fecal coliform bacteria result, the minimum fecal bacteria result average is 4.0 Colonies/100mL, and the maximum is 1624.3 Colonies/100mL. The median result is 108 Colonies/100mL and the mean is 204 Colonies/100mL. The 5 sites with the highest fecal bacteria contamination are Tributary 6 @ Bull Creek (EG), Taylor Slough South @ Reed Park (TSS), Small Toro Canyon Tributary @ Toro Canyon Rd (STC). Blunn Creek @ Little Stacy Park [LISI 4], and Lady Bird Lake @ 1st St [CC].

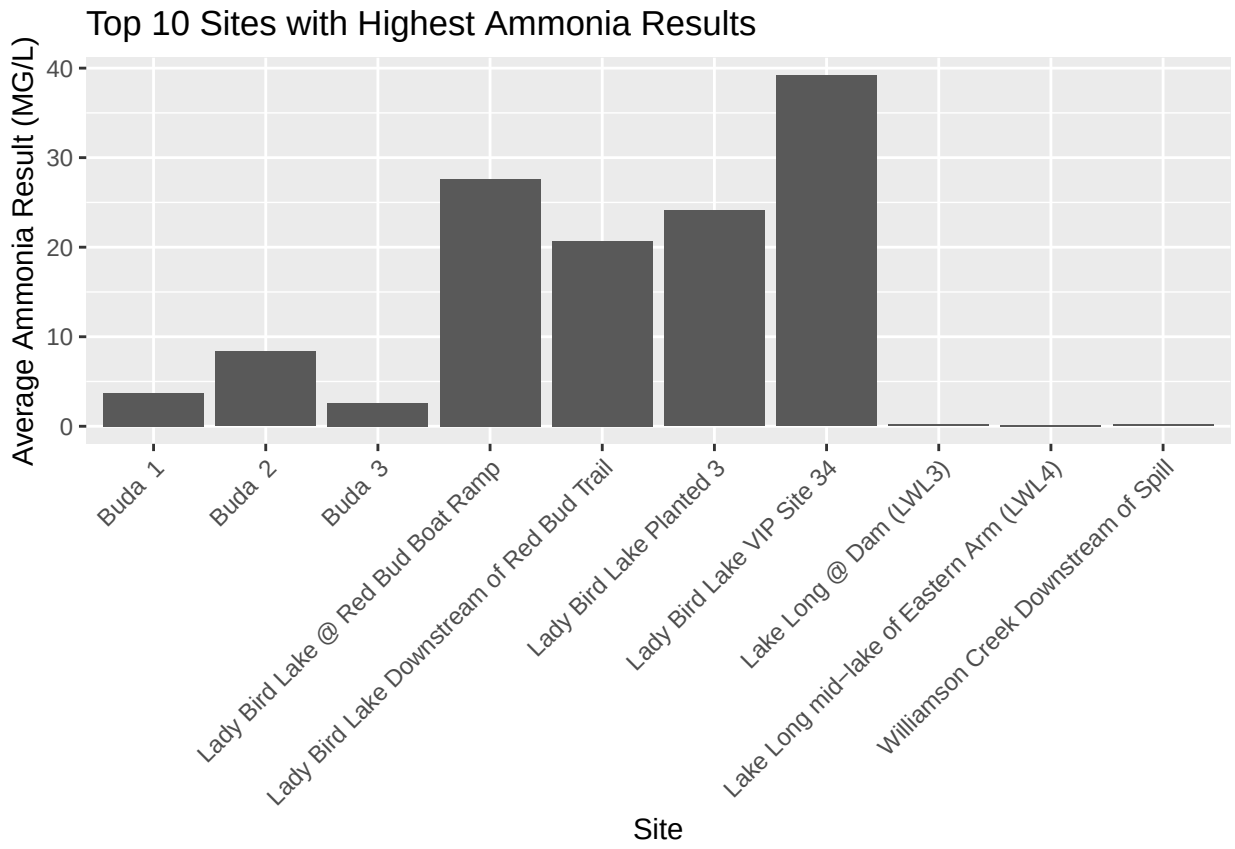
```
samples_clean |>
  group_by(site) |>
  filter(parameter == "AMMONIA AS N")|>
  summarize(ammonia_average = mean(result)) |>
  summary(ammonia_average)
```

```
##      site           ammonia_average
##  Length:102         Min.   : 0.00500
##  Class :character   1st Qu.: 0.02000
##  Mode  :character   Median : 0.02050
##                     Mean   : 1.26923
##                     3rd Qu.: 0.03786
##                     Max.   :39.20000
```

```r
samples_clean |>
  group_by(site) |>
  filter(parameter == "AMMONIA AS N")|>
  summarize(ammonia_average = mean(result)) |>
  slice_max(n = 10, ammonia_average)
```

```
## # A tibble: 10 x 2
##    site                                          ammonia_average
##    <chr>                                                   <dbl>
##  1 Lady Bird Lake VIP Site 34                             39.2
##  2 Lady Bird Lake @ Red Bud Boat Ramp                     27.6
##  3 Lady Bird Lake Planted 3                               24.1
##  4 Lady Bird Lake Downstream of Red Bud Trail             20.7
##  5 Buda  2                                                 8.34
##  6 Buda  1                                                 3.71
##  7 Buda  3                                                 2.62
##  8 Williamson Creek Downstream of Spill                    0.206
##  9 Lake Long @ Dam (LWL3)                                  0.193
## 10 Lake Long mid-lake of Eastern Arm (LWL4)                0.147
```

```r
samples_clean |>
  group_by(site) |>
  filter(parameter == "AMMONIA AS N")|>
  summarize(ammonia_average = mean(result)) |>
  slice_max(n = 10, ammonia_average) |>
  ggplot()+
  geom_col(aes(x = site, y = ammonia_average)) +
  labs(x = "Site", y = "Average Ammonia Result (MG/L)", title = "Top 10 Sites with Highest Ammonia Resul
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

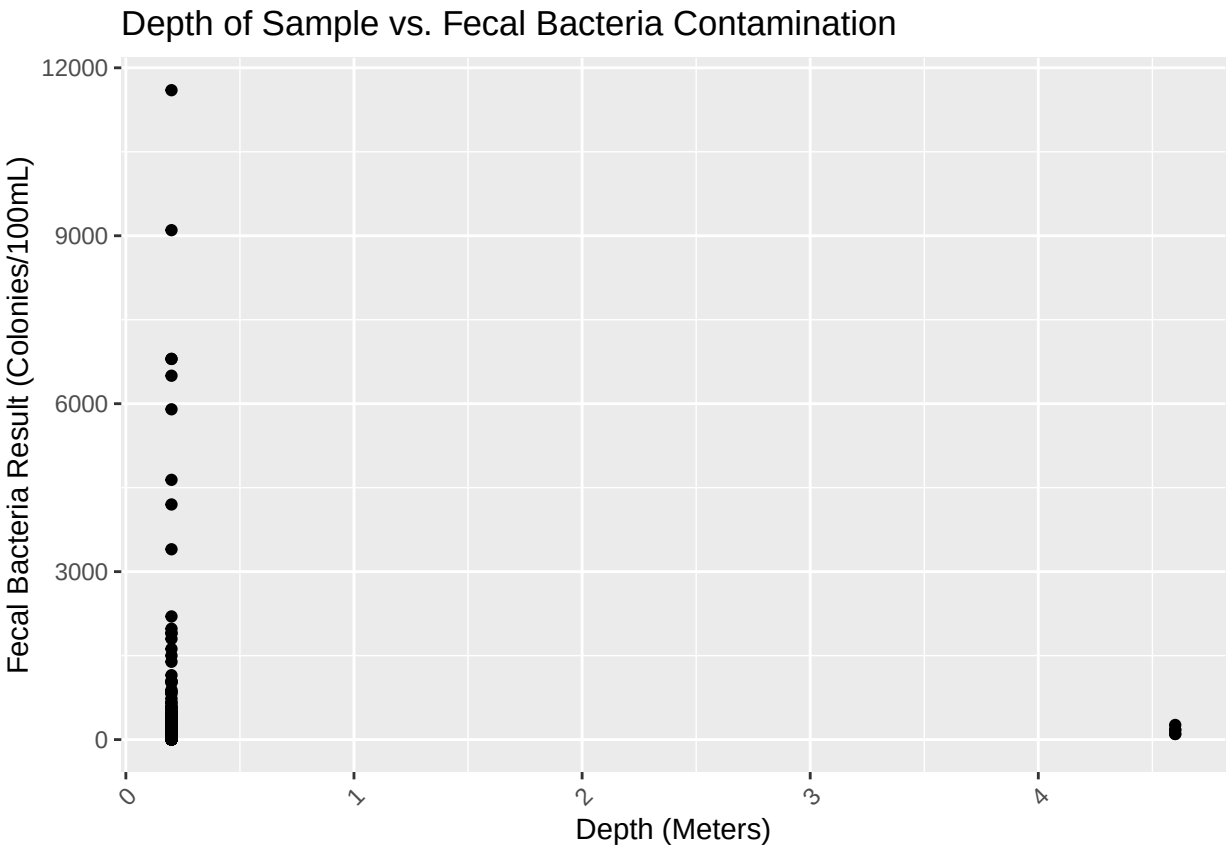


Filtering to observe only the average ammonia result, the minimum ammonia result average is 0.005 MG/L, and the maximum is 39.2 MG/L. The median result is 0.0205 MG/L and the mean is 1.269 MG/L. The 7 sites with the highest average concentration of ammonia are four of the Lady Bird Lake site, 3 of the Buda sites, two of the Lake Long sites, and one of the Williamson Creek sites that was placed downstream of a chemical spill. This high concentration of ammonia across sites implies that on average, some of these water sources, especially Lady Bird Lake and Buda, are highly contaminated with ammonia.

Analyzing Variable Sets: Filter Results vs. Depth

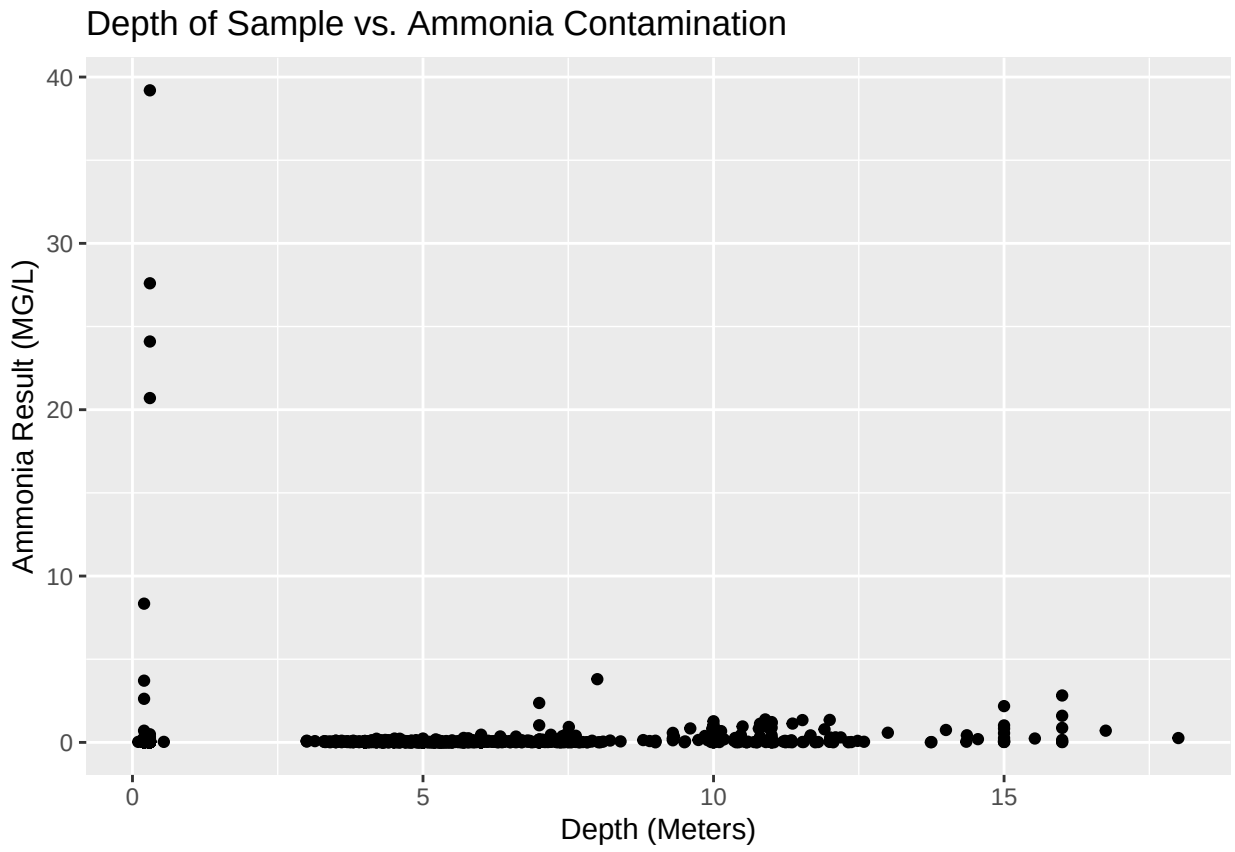
```
samples_clean |>
  filter(parameter == "FECAL COLIFORM BACTERIA")|>
  ggplot()+
  geom_point(aes(x = depth, y = result)) +
  labs(x = "Depth (Meters)", y = "Fecal Bacteria Result (Colonies/100mL)", title = "Depth of Sample vs.
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



Given the visualization, we need a larger variety of depth in sample to be able to appropriately understand the relationship between depth and fecal bacteria contamination result. On average, the results at smaller depths are higher than those at larger depths, but there are also a very large amount of low results at smaller depths.

```
samples_clean |>
  filter(parameter == "AMMONIA AS N")|>
  ggplot()+
  geom_point(aes(x = depth, y = result)) +
  labs(x = "Depth (Meters)", y = "Ammonia Result (MG/L)", title = "Depth of Sample vs. Ammonia Contamina
```



With the larger variation in depth, we can clearly see the relationship between ammonia contamination and depth. Samples with lower depth are much more likely to have higher ammonia contamination, with no ammonia results above 5 MG/L at more than 2.5 meters.

Discussion

The research question focused on how the depth and location affect the filter result of an Austin water sample. Given the large variety of contaminants analyzed in the dataset, I chose to focus on fecal bacteria and ammonia contamination, as trends from these contaminants may apply to other contaminants. When examining filter results versus location, the analysis of fecal coliform bacteria showed that the minimum average result was 4.0 Colonies/100mL, the maximum was 1624.3 Colonies/100mL, the median was 108 Colonies/100mL, and the mean was 204 Colonies/100mL. The highest contamination levels were found at Tributary 6 @ Bull Creek, Taylor Slough South @ Reed Park, Small Toro Canyon Tributary @ Toro Canyon Rd, Blunn Creek @ Little Stacy Park, and Lady Bird Lake @ 1st St. For ammonia, the minimum average result was 0.005 MG/L, the maximum was 39.2 MG/L, the median was 0.0205 MG/L, and the mean was 1.269 MG/L. High concentrations were found at Lady Bird Lake, Buda sites, Lake Long sites, and Williamson Creek downstream of a chemical spill.

In terms of filter results versus depth, higher fecal coliform bacteria contamination results were more common at smaller depths, although there were also many low results at these depths, making the data difficult to make generalizations from. For ammonia, higher contamination was generally more likely at smaller depths, with no results above 5 MG/L at depths greater than 2.5 meters.

My initial expectation was that samples taken around the same location might share filter results and that samples closer to the surface might have fewer contaminants, which was proven correct. However, the strong right skew in depth data and the high contamination levels at certain sites surprised me. Further investigation into why certain sites have higher contamination levels and the impact of specific local factors on water quality would be interesting. I do also want to investigate the locations in context and find reasons they might contain higher levels of specific contaminants, which was difficult to do without map data and the ability to overlay site locations onto a general Austin map.

For the city of Austin, the main takeaway from this exploratory data analysis should be that certain locations in Austin, particularly lakes and specific tributaries, have significantly higher levels of contaminants such as fecal coliform bacteria and ammonia. This indicates a need for targeted water quality improvement efforts in these areas to ensure public health and environmental protection.

The dataset used had many inconsistencies, with 206,950 rows lacking depth information and 145,978 rows missing sample IDs, which were filtered out. There were significant missing values, particularly in the depth and sample ID columns. After cleaning, the dataset was relatively tidy and manageable, but the initial state required substantial filtering and wrangling to make it usable for analysis.

Reflection

One of the biggest challenges I faced working through this project was figuring out how exactly to visualize comparisons between separate variables, and interpreting the results and visualizations accurately. It was especially difficult in areas such as depth, where large amounts of data did not exist, making it difficult to understand trends.

From this process, I learned the importance of data cleaning and preparation. Making sure that the dataset is tidy and free of inconsistencies is an important step that can really impact the quality and accuracy of the analysis. Datasets we may work with in class are often already fairly tidy and easy to work with, but real-life datasets may have large amounts of missing information, or be organized in ways that make them more difficult to work with. In addition, I gained a much deeper understanding of how to interpret and present data effectively. Figuring out how to represent data and create clear and informative visualizations, along with detailed summary statistics, helped in communicating the findings more effectively.

I would like to thank Professor Layla Guyot and TA Walter Wang for their instruction in class and help on this project, as well as dataset owner Robert Clayton, without whom this report would not exist.

References

https://www.jsg.utexas.edu/news/2022/12/saving-austins-water/

https://www.austinwatersolutions.net/understanding-austins-water-quality-issues/

https://data.austintexas.gov/Environment/Water-Quality-Sampling-Data/5tye-7ray/data_preview