Project Check In 2

Who

- Amanda Lee, awl646
- Nabil Yusufali, nay277

Data Set

- https://data.austintexas.gov/Environment/Water-Quality-Sampling-Data/5tye-7ray/about_data (https://data.austintexas.gov/Environment/Water-Quality-Sampling-Data/5tye-7ray/about_data)
- Austin Water Quality Sampling Data
- Each row is a water quality sample by parameter, date, and location in Austin, TX. We are primarily interested in variables watershed, site_type, and result to understand the distribution of water quality across Austin.
- There are 1,475,965 rows and 24 columns in the original data set. Because there is a lot of data, we have limited the data set to bacteria/pathogen samples. The data set ‘water’ contains 34893 rows and 24 columns.

```
# Import 2024 Austin water quality sampling data set
water <- read.csv('https://raw.githubusercontent.com/Amandawlee/sds322_project/refs/heads/main/Water_Quality_Sampling_Data.csv')

# Number of rows and columns in 'water'
dim(water)
```

```
## [1] 34893    24
```

Research Question

- Citizens have a right to know more about their water quality. Especially given that we, as students, use the same water everyday. Therefore, learning more about the water quality could give actionable insights into how to make better water decisions.
- Here are some links that refer to how chemicals can have an effect on water contamination.
    - Texas Regulators Report More Than 250 New Cases of Groundwater Contamination (https://insideclimatenews.org/news/16122024/texas-regulators-report-new-cases-of-groundwater-contamination/)
    - Austin has little to no ‘forever chemicals’ in its drinking water. What did the city do right? (https://www.kut.org/energy-environment/2024-12-06/austin-tx-forever-chemicals-pfas-drinking-water-report)
- Research question(s) to consider:
    - **What are the most common bacteria/pathogens found in bodies of water in Austin?**
    - **Which locations are certain bacteria/pathogens found?**
    - **What time of day are most samples being taken at?**
    - **Which medium contains the most bacteria/pathogens?**
    - **What is the most common sampling method for bacteria/pathogens?**
    - How does the sample date/time of the sample affect the results?
    - How does the sampling method affect the results?
    - How does the sampling medium affect the results?
    - What sampling methods are used for each parameter and parameter type?

Data Wrangling

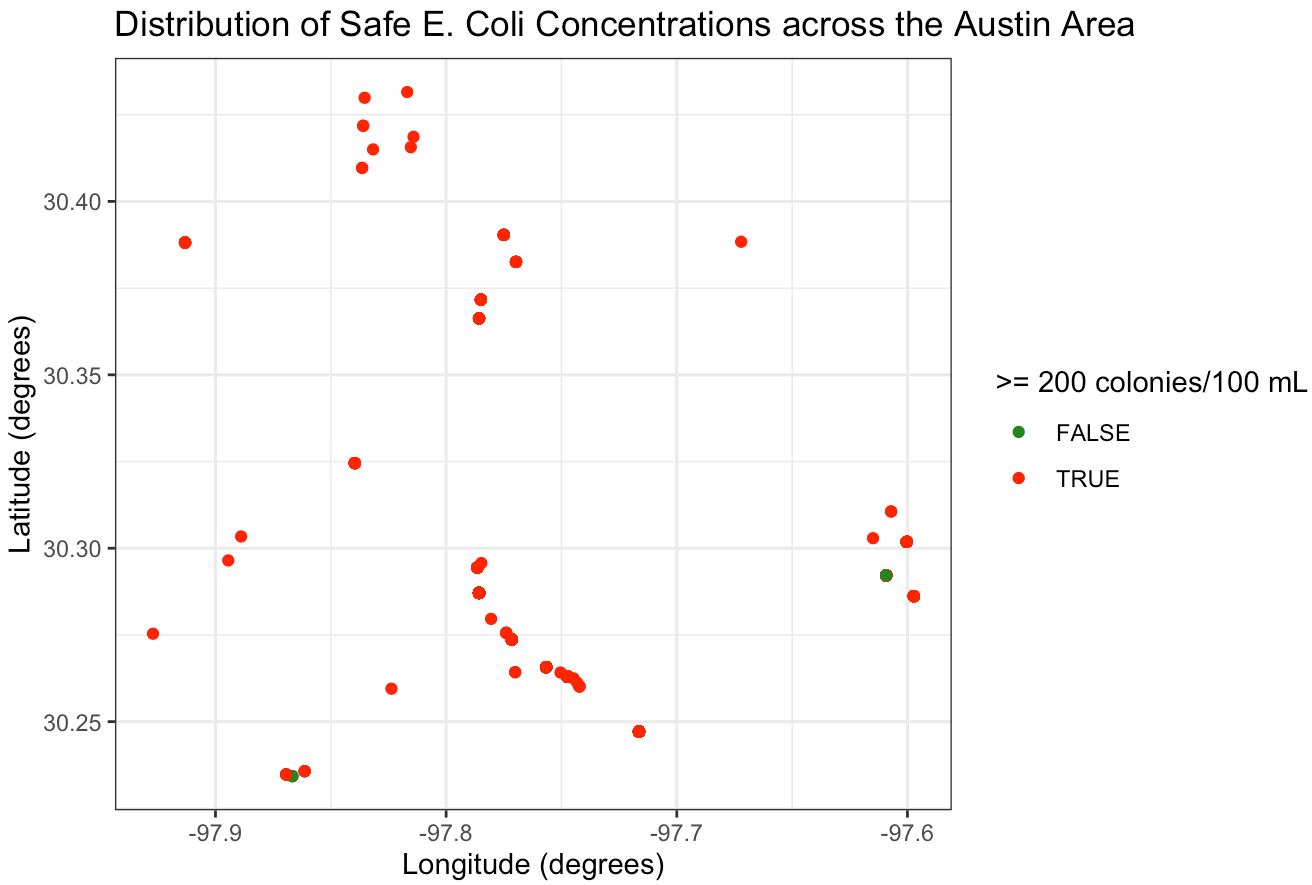
```
# Data tidying/wrangling
water_tidy <- water |>
  # Remove any missing NA values
  na.omit() |>
  # Rename columns to lowercase
  rename_all(~ str_to_lower(.)) |>
  # Select all relevant columns
  select(data_ref_no, lat_dd_wgs84, lon_dd_wgs84, param_type, parameter, result, me
thod) |>
  # Filter for Bacteria/Pathogen parameter type
  filter(param_type == "Bacteria/Pathogens")

head(water_tidy)
```

```
##   data_ref_no lat_dd_wgs84 lon_dd_wgs84         param_type
## 1       26949     30.27152    -97.83134 Bacteria/Pathogens
## 2       26960     30.27508    -97.83621 Bacteria/Pathogens
## 3       26966     30.28423    -97.85214 Bacteria/Pathogens
## 4       30602     30.24492    -98.12569 Bacteria/Pathogens
## 5       30593     30.26364    -97.77303 Bacteria/Pathogens
## 6       69380     30.27152    -97.83134 Bacteria/Pathogens
##                 parameter result    method
## 1 FECAL COLIFORM BACTERIA    127 SM 9221 E
## 2 FECAL COLIFORM BACTERIA     71 SM 9221 E
## 3 FECAL COLIFORM BACTERIA     36 SM 9221 E
## 4 FECAL COLIFORM BACTERIA     88 SM 9222 D
## 5 FECAL COLIFORM BACTERIA     17 SM 9222 D
## 6 FECAL COLIFORM BACTERIA      1 SM 9221 E
```

Exploring One Numeric Variable

```
# Exploring safe E. Coli concentrations in relation to location of samples
water_tidy |>
  filter(parameter == "E COLI BACTERIA") |>
  mutate(ecoli_safe = ifelse(result >= 200, F, T)) |>
  ggplot() +
  #Plot lat/long map and color based on safety
  geom_point(aes(x = lon_dd_wgs84, y = lat_dd_wgs84, color = ecoli_safe)) +
  scale_color_manual(values = c("forestgreen", "red"), name = ">= 200 colonies/100
mL") +
  labs(
    x = 'Longitude (degrees)',
    y = 'Latitude (degrees)',
    title = 'Distribution of Safe E. Coli Concentrations across the Austin Area',
    caption = "Information on safe concentrations of E. Coli in recreational wate
r:\n https://www.knowyourh2o.com/outdoor-4/fecal-coliform-bacteria-in-water"
  ) +
  theme_bw()
```



Information on safe concentrations of E. Coli in recreational water: https://www.knowyourh2o.com/outdoor-4/fecal-coliform-bacteria-in-water

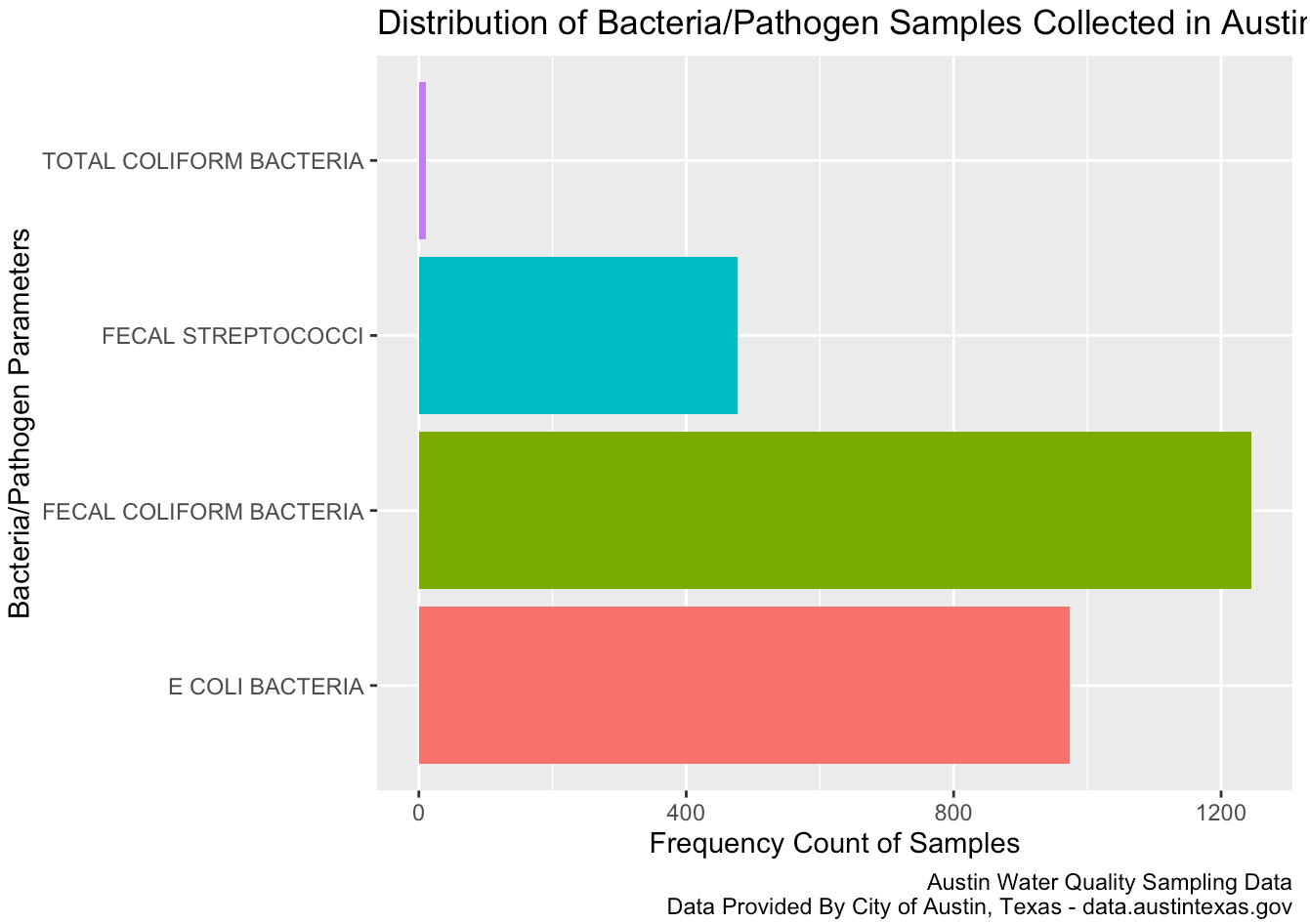
```
water_tidy |>
  filter(parameter == "E COLI BACTERIA") |>
  mutate(ecoli_safe = ifelse(result >= 200, F, T)) |>
  group_by(ecoli_safe) |>
  summarize(count = n())
```

```
## # A tibble: 2 × 2
##   ecoli_safe count
##   <lgl>      <int>
## 1 FALSE         64
## 2 TRUE         909
```

- Other numerical variables to consider:
  - Sample Date: sample_date
  - (Sampling) Result: result
  - Latitude Coordinate: lat_dd_wgs84
  - Longitude Coordinate: lon_dd_wgs84

Exploring One Categorical Variable

```
# Distribution of frequency of each bacteria/pathogen parameter
water_tidy |>
  group_by(parameter) |>
  summarize(count = n(),
            proportion = n()/nrow(water_tidy)) |>
  ggplot(aes(x = count, y = parameter, fill = parameter)) +
  geom_bar(stat = 'identity', show.legend = FALSE) +
  labs(
    title = 'Distribution of Bacteria/Pathogen Samples Collected in Austin',
    x = 'Frequency Count of Samples',
    y = 'Bacteria/Pathogen Parameters',
    caption = 'Austin Water Quality Sampling Data\n Data Provided By City of Austi
n, Texas - data.austintexas.gov')
```




```
# Summary statistics
water_tidy |>
  group_by(parameter) |>
  summarize(count = n(),
            proportion = n()/nrow(water_tidy))
```

```
## # A tibble: 4 × 3
##   parameter               count proportion
##   <chr>                   <int>      <dbl>
## 1 E COLI BACTERIA           973    0.360
## 2 FECAL COLIFORM BACTERIA  1245    0.460
## 3 FECAL STREPTOCOCCI        477    0.176
## 4 TOTAL COLIFORM BACTERIA    11    0.00407
```

- Other categorical variables to consider:
    - Parameter Type: param_type
    - Parameter: parameter
    - Medium: medium
    - Method: method

Questions

- Is it better to try to start with the original data set and tidy it or use the City of Austin Open Data Portal to tidy it first?