Water Research

Cameron Doffing, Isaiah Giebel, Kate Groskreutz, Nguyen Cat Huynh

```
library(readxl)
library(tidyverse)
library(janitor)
OP_21_Water_Use <- read_excel("OP-21_Water_Use_2024.xlsx")
```

Introduction:

Our Website

The Association for the Advancement of Sustainability in Higher Education (AASHE) partners with higher education institutions to promote sustainable practices. The organization's STARS program serves as a concrete metric by which to measure and celebrate schools that reach certain levels of sustainability. The University of St. Thomas is among the 360 institutions that have a valid (non-expired) STARS rating. Given the Office of Sustainability Initiatives (OSI) of St. Thomas wishes to improve their rating, our group decided to explore the performance of participating schools in their ability to reduce water use (called the OP 21 credit). One of the OSI's guiding questions of the data exploration asked for an OP 21 credit score comparison between St. Thomas and one of the two following groups of institutions:

Catholic Benchmark Institutions:

- Creighton University
- Gonzaga University
- Loyola Chicago
- Loyola Marymount
- Santa Clara University
- Seattle University
- University of Dayton
- University of Notre Dame
- University of San Diego

- Villanova University

MN Peer Institutions:

- Bemidji State University
- Carleton College
- College of St. Benedict/St. John's University
- Concordia Moorhead
- Macalester College
- Winona State University
- UMN – Twin Cities
- UMN – Morris
- UMN – Duluth
- Augsburg University
- Concordia in St. Paul
- Hamline University
- St. Kate's University
- St. Olaf College

Initially, our group was provided with an Excel Spreadsheet containing 348 institutions and 58 variables related to water use. However, we needed to first understand how the OP 21 credit score was calculated to understand our variables in context. Following the OP 21 scoring guide (version 2.2), we learned the following:

- Institutions are divided into one of three categories of physical risk quantity, determined by the amount of "water stress and scarcity" as well as "relative water abundance."
- These physical risk quantities determine the total available points that an institution can earn for the OP 21 credit
- The OP 21 score is comprised of three parts:

  1. Reduction in potable water (gallons or cubic meters) use per person
  2. Reduction in potable water use per unit (square feet or square meters) of floor area
  3. Reduction in total water withdrawal per unit (acre or hectare) of vegetated grounds

Comparing the measures of institutions across these three parts (and variables that constituted them) would prove to be primary focus of the advanced visualizations section of the project. However, the data required quite a bit of cleaning to start.

Stage 1: Data Tidying and Wrangling

To begin data cleaning, values in the dataset intended as NA values (but visible as “–” and “**” characters) needed to be recognized as official dataframe NA values.

```r
clean1 <- OP_21_Water_Use |>
  mutate(across(where(is.character), ~ na_if(.x,  "--"))) |>
  mutate(across(where(is.character), ~ na_if(.x,  "**")))
```

Next, four variables of dates, which were previously stored as serial date numbers (in the tens of thousands), needed to be converted to the Date type.

```r
clean1[14:17] <- lapply(clean1[14:17], function(x) as.Date(as.numeric(as.character(x)),
    origin = "1899-12-30"))
```

A general cleaning of names soon followed using the `janitor` package, with the sixth column receiving a new name to replace the previous verbose title which contained special characters difficult to type.

```r
colnames(clean1)[6] <- "Physical Risk Quantity"
clean1 <- clean1 |>
  clean_names()
```

Similarly, the original excel spreadsheet resolved the non-English letters of seven university names to incorrect characters such as the square root and copyright symbol, so these names were manually fixed as well.

```r
clean1[c(92, 161,223:224, 320:324,348), "institution"] <- c("HEC Montréal",
"Polytechnique Montréal", "Universidad Autónoma de Tamaulipas",
"Universidad Científica del Sur", "Université Laval",
"Université Téluq", "Université de Montréal",
"Université de Sherbrooke",
"Université du Québec à Montréal",
"École de Technologie Supérieure")
```

Many of the variables needed to be changed from type character to numeric.

```
clean1[c(7, 8, 9, 11:12, 20:ncol(clean1))] <-
  lapply(clean1[c(7, 8, 9, 11:12, 20:ncol(clean1))], as.numeric)
```

Two of the schools of interest for comparison (University of Minnesota-Duluth and Gonzaga University) were not present in the original data set despite having STARS ratings under version 2.2. Therefore, these two institutions were added manually.

```
new_rows <- tibble(
  institution = c("University of Minnesota, Duluth", "Gonzaga University"),
  date_submitted = as.Date(c("2024-09-11", "2020-09-23")),
  version = c(2.2, 2.2),
  status = c("Pursuing", "Pursuing"),
  physical_risk_quantity = c("Low to Medium", "Low"),
  total_water_withdrawal_performance_year = c(111364484, 116373840),
  total_water_withdrawal_baseline_year = c(143723712, 201530682),
  potable_water_use_performance_year = c(61513276, 116373840),
  potable_water_use_baseline_year = c(74541192, 201530682),
  start_date_performance_year_or_3_year_period = as.Date(c("2021-07-01", "2018-06-01")),
  end_date_performance_year_or_3_year_period = as.Date(c("2022-06-30", "2019-05-31")),
  start_date_baseline_year_or_3_year_period = as.Date(c("2007-01-01", "2008-06-01")),
  end_date_baseline_year_or_3_year_period = as.Date(c("2007-12-31", "2009-05-31")),
  a_brief_description_of_when_and_why_the_water_use_baseline_was_adopted = c(
    "We aligned water baseline with our first greenhouse gas inventory baseline.",
    "2009 was the year we started doing the GHG calculator for our CAP."
  ),
  number_of_students_resident_on_site_performance_year = c(2686, 2776),
  number_of_students_resident_on_site_baseline_year = c(2671, 2696),
  number_of_employees_resident_on_site_performance_year = c(6, 47),
  number_of_employees_resident_on_site_baseline_year = c(0, 45),
  number_of_other_individuals_resident_on_site_performance_year = c(4, 0),
  number_of_other_individuals_resident_on_site_baseline_year = c(0, 0),
  total_full_time_equivalent_student_enrollment_performance_year = c(9166, 7364),
  total_full_time_equivalent_student_enrollment_baseline_year = c(11184, 6823),
  full_time_equivalent_of_employees_performance_year = c(1408, 1440),
  full_time_equivalent_of_employees_baseline_year = c(1506, 1060),
  full_time_equivalent_of_students_enrolled_exclusively_in_distance_education_performance_year
  full_time_equivalent_of_students_enrolled_exclusively_in_distance_education_baseline_year = c
  weighted_campus_users_performance_year = c(8486, 6311.25),
  weighted_campus_users_baseline_year = c(10185.25, 6176),
  potable_water_use_per_weighted_campus_user_performance_year = c(7248.80, 18439.11),
  potable_water_use_per_weighted_campus_user_baseline_year = c(7318.54, 32631.26),
  percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline = c(0.0095,
  gross_floor_area_of_building_space_performance_year = c(3298129, 3060418),
  gross_floor_area_of_building_space_baseline_year = c(3123397, 2378463),
  potable_water_use_per_unit_of_floor_area_performance_year = c(18.65, 38.03),
  potable_water_use_per_unit_of_floor_area_baseline_year = c(23.87, 84.73),
  percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline = c(.2185, .55
  area_of_vegetated_grounds_performance_year = c(289.50, 66),
  area_of_vegetated_grounds_baseline_year = c(291, 66),
  total_water_withdrawal_per_unit_of_vegetated_grounds_performance_year = c(384678.70, 1763240
  total_water_use_per_unit_of_vegetated_grounds_baseline_year = c(493895.92, 3053495.18),
  percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baseline =
)

clean1 <- clean1 |> bind_rows(new_rows)
```

As outlined on the OP 21 scoring guide (version 2.2), schools were assigned one of three point maximums (4/3, 5/3, or 2) for their three-part score. This assignment was determined by the schools' physical risk quantities, as replicated in our wrangling with the creation of a `max_part_pts` column.

```
clean1 <- clean1 |>
  mutate(
    max_part_pts = case_when(
      physical_risk_quantity == "Low" ~ 1 + 1/3,
      physical_risk_quantity == "Low to Medium" ~ 1 + 1/3,
      physical_risk_quantity == "Medium to High" ~ 1 + 2/3,
      physical_risk_quantity == "High" ~ 2,
      physical_risk_quantity == "Extremely High" ~ 2,
      TRUE ~ NA_real_
    )
  )
```

Using the `max_part_pts` column and following the part 1-3 points earned formulas on the scoring guide, the final OP 21 score for each school was calculated. Although not mentioned in the scoring guide but required to recreate the original values, schools scoring below 0 or above the maximum in a part had their score rounded to the respective bound (0 or max, whichever is closer).

```
clean1 <- clean1 |>
  mutate(p1_pts = pmin(pmax(0,
as.numeric((clean1[[59]]/0.3)*((clean1[[36]]-clean1[[35]])/clean1[[36]]))), clean1[[59]]),

p2_pts = pmin(pmax(0,
as.numeric((clean1[[59]]/0.3)*((clean1[[43]]-clean1[[42]])/clean1[[43]]))), clean1[[59]]),

p3_pts = pmin(pmax(0,
as.numeric((clean1[[59]]/0.3)*((clean1[[50]]-clean1[[49]])/clean1[[50]]))), clean1[[59]]),

         op21_score = p1_pts+p2_pts+p3_pts
         )
```

Finally, a categorical variable `risk_group`, dividing the schools among the same logic as the `max_part_pts` variable, was created to aid in summary statistics and visualizations wishing to compare schools by a condensed (three groups instead of five) categorical representation of their physical risk quantity.

```
clean1 <- clean1 |>
  mutate(
    risk_group = case_when(
      physical_risk_quantity == "Low" ~ "1 (Low and Low to Medium)",
      physical_risk_quantity == "Low to Medium" ~ "1 (Low and Low to Medium)",
      physical_risk_quantity == "Medium to High" ~ "2 (Medium to High)",
      physical_risk_quantity == "High" ~ "3 (High and Extremely High)",
      physical_risk_quantity == "Extremely High" ~ "3 (High and Extremely High)",
      TRUE ~ NA_character_
    )
  )
```

To better compare institutions across different physical risk quantities, a op_21_percent variable was created to represent the credit score as a percent rather than a fraction.

```
clean1 <- clean1 |>
  mutate(op_21_percent = clean1[[63]] / (clean1[[59]]*3))
```

Stage 2: Summary Statistics

For calculating summary statistics, the variables concerning OP 21 score, potable water use per weighted campus user reduction (part 1 points), potable water use per gross square meter/foot of floor area reduction (part 2 points), and total water use per acre of vegetated ground (part 3 points) were of most interest.

When separating these four variables by risk group, a few common characteristics are revealed. First, at 212 schools, risk group 1 (low and low to medium risk schools) contains over 100 more schools than risk group 2 (medium to high risk) and risk group 3 (high and extremely high risk) combined, at 57 and 51 schools respectively. However, risk group 3 consistently has the highest measures of standard deviation across the four variables of the three risk groups.

Eleven of the twelve summary statistics are skewed to the left because the medians are greater than the means. However, for the distribution of part 3 points for risk group 2 schools, the distribution is slightly skewed to the right, as the mean exceeds the median by 0.04 points.

```r
clean1 |>
  group_by(risk_group) |>
  summarize(
    min = min(op21_score, na.rm = TRUE),
    Q1 = quantile(op21_score, probs = 0.25, na.rm = TRUE),
    median = median(op21_score, na.rm = TRUE),
    Q3 = quantile(op21_score, probs = 0.75, na.rm = TRUE),
    max = max(op21_score, na.rm = TRUE),
    mean = mean(op21_score, na.rm = TRUE),
    sd = sd(op21_score, na.rm = TRUE),
    IQR = IQR(op21_score, na.rm = TRUE),
    count = n()
  ) |>
  slice(-4)
```

```
# A tibble: 3 x 10
  risk_group                  min    Q1 median    Q3   max  mean    sd   IQR count
  <chr>                     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
1 1 (Low and Low to Medi~       0  1.33   2.67  4        4  2.49  1.41  2.67   214
2 2 (Medium to High)            0  1.58   3.52  4.80     5  3.11  1.73  3.22    57
3 3 (High and Extremely ~       0  2.23   4.90  6        6  4.04  2.07  3.77    51
```

```r
clean1 |>
  group_by(risk_group) |>
  summarize(
    min = min(p1_pts, na.rm = TRUE),
    Q1 = quantile(p1_pts, probs = 0.25, na.rm = TRUE),
    median = median(p1_pts, na.rm = TRUE),
    Q3 = quantile(p1_pts, probs = 0.75, na.rm = TRUE),
    max = max(p1_pts, na.rm = TRUE),
    mean = mean(p1_pts, na.rm = TRUE),
    sd = sd(p1_pts, na.rm = TRUE),
    IQR = IQR(p1_pts, na.rm = TRUE),
    count = n()
  ) |>
  slice(-4)
```

```
# A tibble: 3 x 10
  risk_group              min      Q1 median    Q3   max  mean    sd   IQR count
  <chr>                 <dbl>   <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
1 1 (Low and Low to Me~     0 0.00207  0.941  1.33  1.33 0.760 0.564  1.33   214
2 2 (Medium to High)        0 0.572    1.35   1.67  1.67 1.11  0.668  1.09    57
3 3 (High and Extremel~     0 0.720    2      2     2    1.37  0.825  1.28    51
```

```r
clean1 |>
  group_by(risk_group) |>
  summarize(
    min = min(p2_pts, na.rm = TRUE),
    Q1 = quantile(p2_pts, probs = 0.25, na.rm = TRUE),
    median = median(p2_pts, na.rm = TRUE),
    Q3 = quantile(p2_pts, probs = 0.75, na.rm = TRUE),
    max = max(p2_pts, na.rm = TRUE),
    mean = mean(p2_pts, na.rm = TRUE),
    sd = sd(p2_pts, na.rm = TRUE),
    IQR = IQR(p2_pts, na.rm = TRUE),
    count = n()
  ) |>
  slice(-4)
```

```
# A tibble: 3 x 10
```

```
  risk_group                min    Q1 median    Q3   max  mean    sd   IQR count
  <chr>                   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
1 1 (Low and Low to Medi~     0 0.621   1.33  1.33  1.33 0.962 0.507 0.713   214
2 2 (Medium to High)          0 0.745   1.67  1.67  1.67 1.22  0.636 0.922    57
3 3 (High and Extremely ~     0 1.12    2     2     2     1.52  0.747 0.877    51
```

```r
clean1 |>
  group_by(risk_group) |>
  summarize(
    min = min(p3_pts, na.rm = TRUE),
    Q1 = quantile(p3_pts, probs = 0.25, na.rm = TRUE),
    median = median(p3_pts, na.rm = TRUE),
    Q3 = quantile(p3_pts, probs = 0.75, na.rm = TRUE),
    max = max(p3_pts, na.rm = TRUE),
    mean = mean(p3_pts, na.rm = TRUE),
    sd = sd(p3_pts, na.rm = TRUE),
    IQR = IQR(p3_pts, na.rm = TRUE),
    count = n()
  ) |>
  slice(-4)
```

```
# A tibble: 3 x 10
  risk_group                min    Q1 median    Q3   max  mean    sd   IQR count
  <chr>                   <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <int>
1 1 (Low and Low to Medi~     0 0.114  0.877  1.33  1.33 0.769 0.559  1.22   214
2 2 (Medium to High)          0 0      0.745  1.67  1.67 0.788 0.678  1.67    57
3 3 (High and Extremely ~     0 0.159  1.49   2     2     1.15  0.851  1.84    51
```

For the fifth variable, exploring total water use (or withdrawal) of the performance years, separated by risk group, was chosen. This variable may be useful in understanding why schools of similar size or physical risk quantity score differently in various parts of the overall OP 21 score, an investigation of interest to the University of St. Thomas. Interestingly, the distribution of the risk group 1 has a standard deviation that greatly surpasses that of the other two groups, most likely due to max and min values that are larger and smaller (respectively) than risk groups 2 and 3. The IQR, a measure of spread more resistant to outliers, may be a better metric to compare variation between the risk groups for further analysis.

```
clean1 |>
  group_by(risk_group) |>
  summarize(
    min = min(total_water_withdrawal_performance_year, na.rm = TRUE),
    Q1 = quantile(total_water_withdrawal_performance_year, probs = 0.25, na.rm = TRUE),
    median = median(total_water_withdrawal_performance_year, na.rm = TRUE),
    Q3 = quantile(total_water_withdrawal_performance_year, probs = 0.75, na.rm = TRUE),
    max = max(total_water_withdrawal_performance_year, na.rm = TRUE),
    mean = mean(total_water_withdrawal_performance_year, na.rm = TRUE),
    sd = sd(total_water_withdrawal_performance_year, na.rm = TRUE),
    IQR = IQR(total_water_withdrawal_performance_year, na.rm = TRUE),
    count = n()
  ) |>
  slice(-4)
```

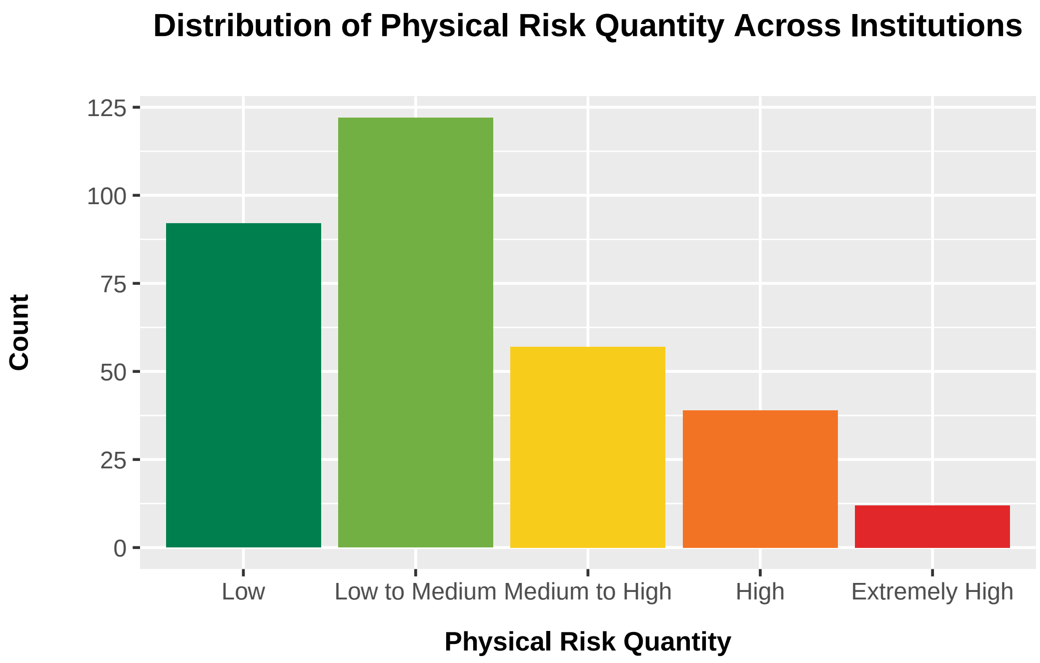
```
# A tibble: 3 x 10
  risk_group          min      Q1 median      Q3      max    mean       sd     IQR count
  <chr>             <dbl>   <dbl>  <dbl>   <dbl>    <dbl>   <dbl>    <dbl>   <dbl> <int>
1 1 (Low and Lo~ 5.70e4 2.81e7 6.97e7 1.88e8 6.98e11 3.72e9 4.78e10 1.60e8   214
2 2 (Medium to ~ 1.45e6 3.47e7 9.54e7 2.17e8 8.98e 8 1.89e8 2.23e 8 1.83e8    57
3 3 (High and E~ 3.19e6 5.61e7 1.23e8 2.75e8 1.10e 9 2.40e8 2.76e 8 2.19e8    51
```

Identifying the three parts of the OP 21 score as measuring percentage reduction in water use across different situations, the creation of visualizations comparing school performance of these three parts was of interest. However, two schools were missing values for the percentage reduction despite having OP 21 credit scores, so these percentages were recalculated.

```
clean1 <- clean1 |>
  mutate(percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline = (cl
         percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline = ((cl
         )
```

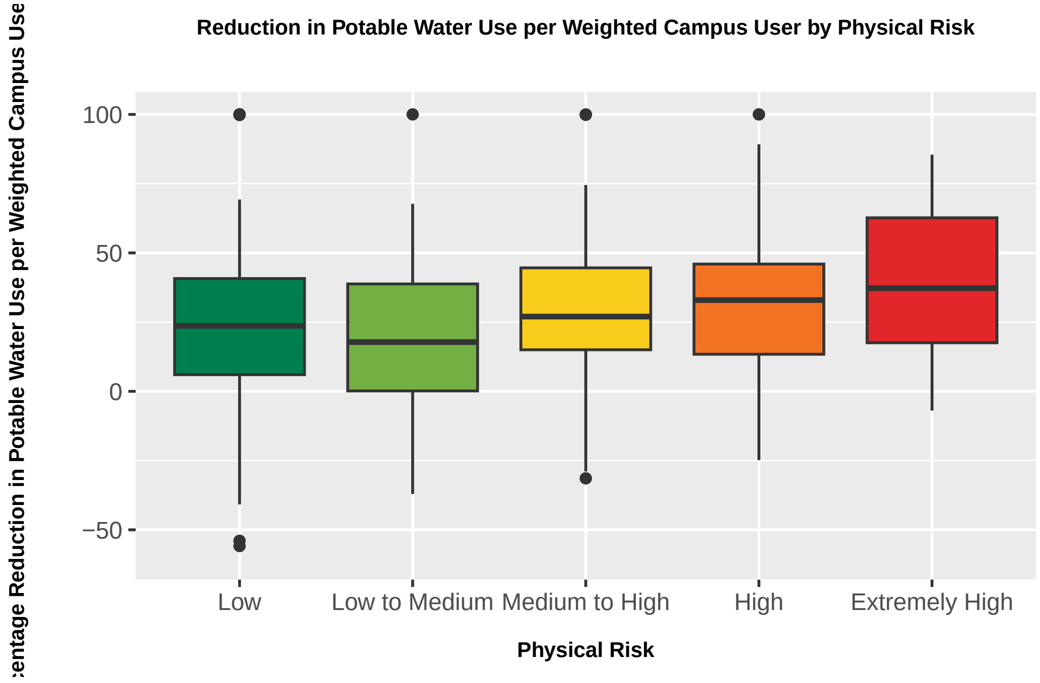
Stage 3: Data Visualizations

```
order <- c("Low", "Low to Medium", "Medium to High", "High", "Extremely High")

physrisk <- clean1 |>
  filter(!is.na(physical_risk_quantity)) |>
  ggplot(aes(x =
factor(physical_risk_quantity, levels = order),
fill = physical_risk_quantity)) +
  scale_fill_manual(values = c("Low" = "#007f4e",
                               "Low to Medium" = "#72b043",
                               "Medium to High" = "#f8cc1b",
                               "High" = "#f37324",
                               "Extremely High" = "#e12729")) +
  geom_bar(stat = "count", na.rm = TRUE) +
  labs(title = "Distribution of Physical Risk Quantity Across Institutions",
x = "Physical Risk Quantity", y = "Count", fill = "Risk Level") +
  theme(
    plot.title =
element_text(size = 12, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y =
    element_text(size = 10, face = "bold", margin = margin(r = 20)),
    axis.title.x =
    element_text(size = 10, face = "bold", margin = margin(t = 10)),
    legend.position = "none")
physrisk
```



Here, we are looking at the distribution of physical risk quantity across each institution in our dataset. We created a bar chart that counts the number of institutions in each classification of physical risk quantity. I filtered to include each institution that had a value for their physical risk quantity. To get the classifications in the correct order, I had to manually write an order that would make the more sense. I then imported a specific color palette that would be helpful for a viewer to visualize the context of the graph. From this visualization, we can see that a majority of these institutions are classified as having a physical risk quantity of “Low” or “Low to Medium”. As we move along the x-axis, we see the counts of institutions in each classification decreasing as the physical risk quantity gets more extreme, which is a good sign.

```
pctreduc <- clean1 |>
filter(
!is.na(
percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline)) |>
  mutate(percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline=perce

pctreduc |>
  ggplot(aes(x = factor(physical_risk_quantity, levels = order),
y = percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline,
fill = physical_risk_quantity)) +
  scale_fill_manual(values = c("Low" = "#007f4e",
                               "Low to Medium" = "#72b043",
                               "Medium to High" = "#f8cc1b",
                               "High" = "#f37324",
                               "Extremely High" = "#e12729")) +
  geom_boxplot() +
  labs(
title = str_wrap(
"Reduction in Potable Water Use per Weighted Campus User by Physical Risk"),
x = "Physical Risk",
y = str_wrap("Percentage Reduction in Potable Water Use per Weighted Campus User (%)")) +
  theme(
    plot.title = element_text(size = 8, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),
    legend.position = "none"
  ) +
  scale_y_continuous(limits = c(-60, 100))
```

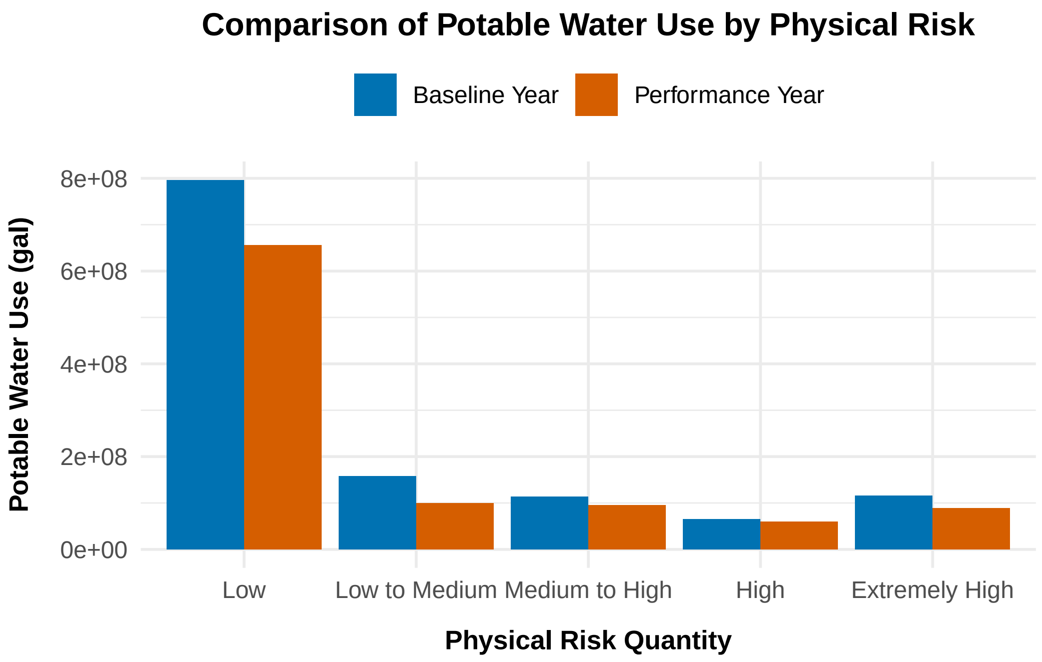


For this visualization, we are looking at the percentage of reduction in potable water use per weighted campus user from the baseline amount. We formatted this as a side-by-side boxplot, where we can visualize the differences between each classification of physical risk quantity. We used the same order for the x-axis as the visualization above, along with their corresponding colors. When we investigate this visualization, we can see that on average, the institutions with the better classifications have a lower percentage reduction in potable water use per weighted campus user from the baseline amount. This would make sense because if an institution is doing really well in their physical risk classification, it would be difficult to reduce when you are already really low. As we move to the right along the x-axis, we see that on average, the classification groups with higher risk have a higher reduction in potable water use per weighted campus user from the baseline amount. Contextually, we can understand that this would make sense because these institutions have a lot more room for improvement in their reduction. We want these groups especially to have a higher reduction, since they have a higher physical risk quantity.

Recognizing potable water use to be a common variable in the calculations of a school's part 1 and part 2 scores, a data visualization comparing total potable water use between baseline and performance years was of interest. Given the schools of focus are within two groups (Catholic Benchmark and MN Peer institutions), The dataset was filtered to just include these two groups, utilizing `pivot_longer` to add columns specifiying if a school's potable water use data corresponded to their performance or baseline year.

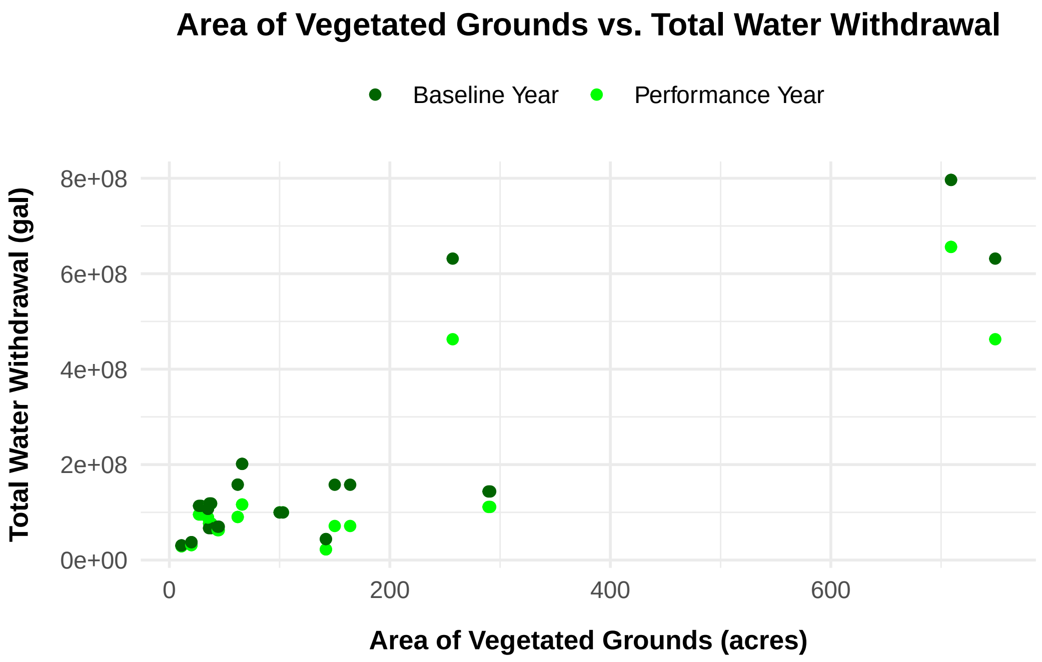
```r
catholic_benchmark_institutions <- c(
  "Creighton University",
  "Gonzaga University",
  "Loyola University Chicago",
  "Loyola Marymount University",
  "Santa Clara University",
  "Seattle University",
  "University of Dayton",
  "University of Notre Dame",
  "University of San Diego",
  "Villanova University",
  "University of St. Thomas"
)

mn_peer_institutions <- c(
  "Bemidji State University", "
  Carleton College",
  "College of Saint Benedict",
  "St. John's University",
  "Concordia College - Moorhead",
  "Macalester College",
  "Winona State University",
  "University of Minnesota, Twin Cities",
  "University of Minnesota, Morris",
  "University of Minnesota, Duluth",
  "Augsburg University",
  "Concordia in St. Paul",
  "Hamline University",
  "St. Kate's University",
  "St. Olaf College",
  "University of St. Thomas"
```

```
)

data_long <- clean1 |>
  filter(!is.na(potable_water_use_performance_year) &
           !is.na(potable_water_use_baseline_year)) |>
  pivot_longer(
    cols = c(
      potable_water_use_performance_year,
      potable_water_use_baseline_year
    ),
    names_to = "year",
    values_to = "Potable_Water_Use"
  ) |>
  mutate(
    year = recode(year,
                  potable_water_use_performance_year = "Performance Year",
                  potable_water_use_baseline_year = "Baseline Year")
  ) |>
  filter(institution %in% catholic_benchmark_institutions |
         institution %in% mn_peer_institutions)

# Create the plot
ggplot(data_long, aes(x = factor(physical_risk_quantity, levels = order),
y = Potable_Water_Use, fill = year)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(
    values =
      c("Baseline Year" = "#0072B2", "Performance Year" = "#D55E00")) +
  labs(
    title = "Comparison of Potable Water Use by Physical Risk",
    x = "Physical Risk Quantity",
    y = "Potable Water Use (gal)",
    fill = "Year"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
    axis.title.y = element_text(size = 10, face = "bold", margin = margin(r = 10)),
   axis.title.x = element_text(size = 10, face = "bold", margin = margin(t = 10)),
   legend.position = "top"
  ) +
  guides(fill = guide_legend(title = NULL))
```



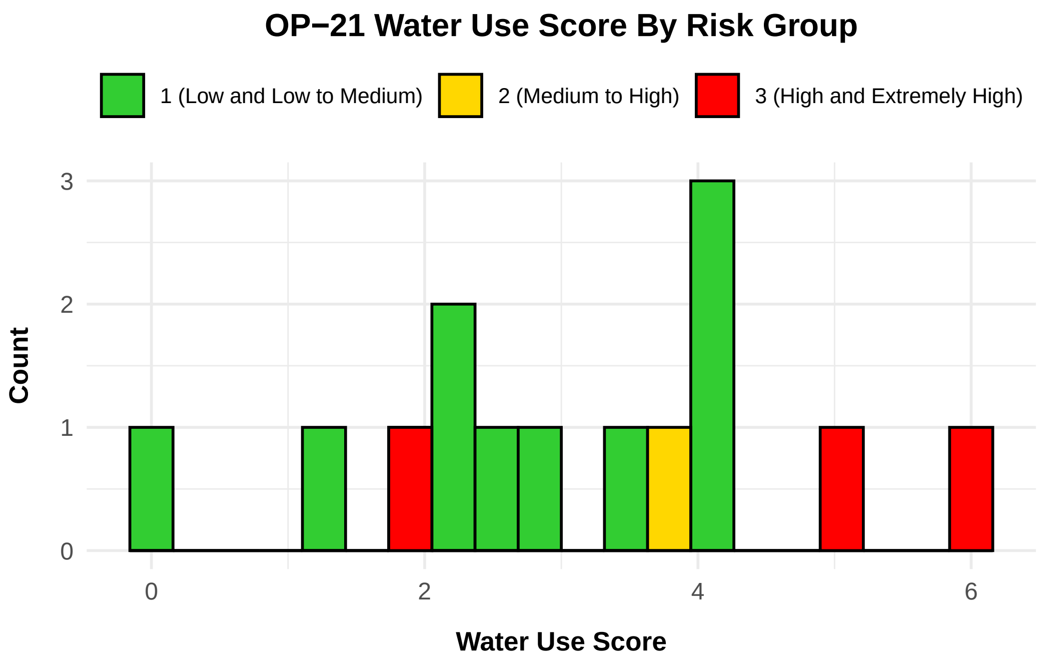
The resulting side-by-side bar chart reveals that among these two groups of schools, their combined potable water use decreases in the performance year as compared to the baseline year. This reduction makes sense, as these AASHE participants strive to increase their sustainability, which an increase in potable water use would detract from. The apparent higher use of potable water in the low risk column is representative of the greater number of schools of risk group 1 in the filtered dataset.

```
clean1_transformed <- clean1 |>
  select(-area_of_vegetated_grounds) |>
  filter(institution %in% catholic_benchmark_institutions |
         institution %in% mn_peer_institutions) |>
  pivot_longer(
    cols =
c(total_water_withdrawal_performance_year,
  total_water_withdrawal_baseline_year),
    names_to = "year",
    values_to = "total_water_withdrawal") |>
  pivot_longer(
    cols =
c(area_of_vegetated_grounds_performance_year,
  area_of_vegetated_grounds_baseline_year),
    names_to = "placeholder",
    values_to = "area_of_vegetated_grounds") |>
  mutate(
    year_type = recode(year,
    total_water_withdrawal_performance_year = "Performance Year",
    total_water_withdrawal_baseline_year = "Baseline Year")
  )

ggplot(clean1_transformed, aes(x = area_of_vegetated_grounds,
                               y = total_water_withdrawal)) +
  geom_point(aes(color = year_type)) +
  labs(
    title = "Area of Vegetated Grounds vs. Total Water Withdrawal",
    x = "Area of Vegetated Grounds (acres)",
    y = "Total Water Withdrawal (gal)"
  ) +
  guides(color = guide_legend(title = NULL)) +
  scale_color_manual(
  values =
  c("Baseline Year" = "darkgreen", "Performance Year" = "green")) +
  theme_minimal() +
  theme(
plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
axis.title.y = element_text(size = 10, face = "bold", margin = margin(r = 10)),
 axis.title.x = element_text(size = 10, face = "bold", margin = margin(t = 10)),
 legend.position = "top"
) +
guides(fill = guide_legend(title = NULL))
```



This scatter plot shows the area of vegetated grounds in acres versus the total water withdrawal for each institution in each of the two institution categories provided for the project. The color of the dots indicates whether the measurements are for the performance year or the baseline year. To further prepare the data for use in this graphic we needed to filter on the desired schools and perform to pivot longers. The pivot longers are what allowed us to differentiate year type by color in the graph by breaking both total water withdrawal and area of vegetated grounds into their year types (baseline or performance). There are a few patterns/conclusions we can find from the graph. Almost all schools have a very apparent drop in total water withdrawal while keeping essentially the same area of vegetated grounds. There is a weak positive correlation between these two variables. Most institutions are clustered in the bottom left corner of the graph which indicates lower values for both variables, but there are a couple higher values for both variables across the plot that could be considered outliers.

```
clean2 <- clean1 |>
    filter(institution %in% catholic_benchmark_institutions |
institution %in% mn_peer_institutions)

ggplot(clean2, aes(x = op21_score)) +
  geom_histogram(
aes(fill = risk_group),
bins = 20,
position = "identity", color = "black") +
  labs(
    title = "OP-21 Water Use Score By Risk Group",
    x = "Water Use Score",
    y = "Count",
    fill = "Risk Group"
  ) +
  scale_fill_manual(
values = c("1 (Low and Low to Medium)" = "limegreen",
"2 (Medium to High)" = "gold",
"3 (High and Extremely High)" = "red")) +
  theme_minimal() +
  theme(
    plot.title = element_text(
  size = 12, face = "bold", hjust = 0.5),
    axis.title.y = element_text(
  size = 10, face = "bold", margin = margin(r = 10)),
    axis.title.x = element_text(
  size = 10, face = "bold", margin = margin(t = 10)),
    legend.position = "top",
    legend.text = element_text(size = 8)
  ) +
  guides(fill = guide_legend(title = NULL))
```



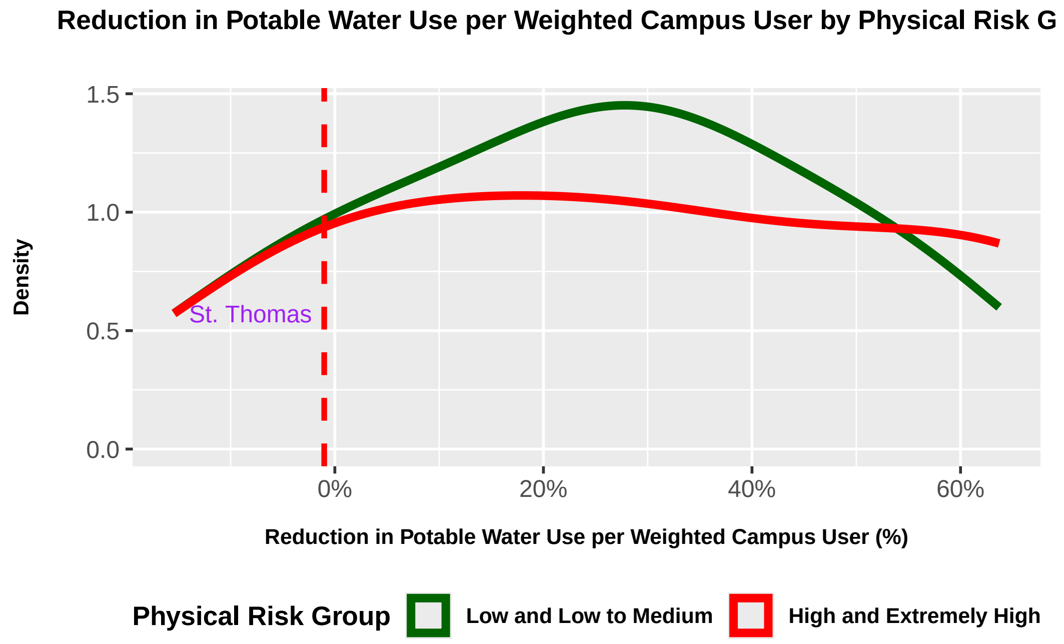
This histogram shows the distribution of OP-21 water use scores with coloring for the risk group. It should be noted that the highest score a school with a risk group of 1 can achieve is 4, for risk group 2 the highest is a 5, and for risk group 3 the highest is a 6. The distribution is somewhat normal but far from perfect. It is hard to have a normal distribution with only 14 cases. The values are spread across from 0 to 6 and there are no outliers.

Phase 2: Advanced Visualizations

The first three visualizations concern just Catholic Benchmark Institutions, observing where St. Thomas lies on density plots of the three percentage reduction variables of interest. Through visual inspection, we can determine where St. Thomas' OP 21 performance lies on the distribution of Catholic Benchmark Institutions.

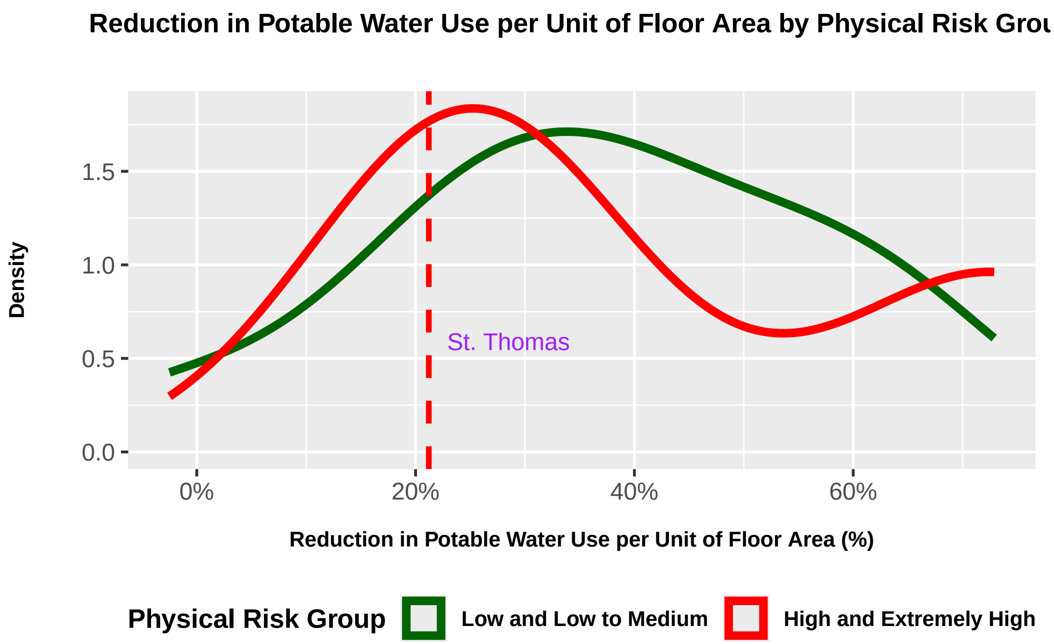
```
universities_of_interest <- c(
  "Creighton University",
  "Gonzaga University",
  "Loyola University Chicago",
  "Loyola Marymount University",
  "Santa Clara University",
  "Seattle University",
  "University of Dayton",
  "University of Notre Dame",
  "University of San Diego",
  "Villanova University",
  "University of St. Thomas"
)

catholic_unis <- clean1 %>%
  filter(institution %in% universities_of_interest)
```

```
ggplot(data = clean2, aes(x = percentage_reduction_in_potable_water_use_per_weighted_campus_use
  geom_density(size = 1.5) +
  scale_color_manual(
    values = c('1 (Low and Low to Medium)' = 'darkgreen', '3 (High and Extremely High)' = 'red
    breaks = c('1 (Low and Low to Medium)', '3 (High and Extremely High)'),
    labels = c('Low and Low to Medium', 'High and Extremely High')
  ) +
  labs(
    color = "Physical Risk Group",
    x = "Reduction in Potable Water Use per Weighted Campus User (%)",
    y = "Density",
    title = "Reduction in Potable Water Use per Weighted Campus User by Physical Risk Group"
  ) +
  scale_x_continuous(labels = scales::percent) +
  geom_vline(xintercept = -0.010209659, linetype = "dashed", color = "red", size = 1) +
  annotate("text", x = -0.010209659, y = 0.5, label = "St. Thomas",
           color = "purple", vjust = -0.5, hjust = 1.1, size = 3) +
  theme(
    plot.title = element_text(size = 10, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),

    legend.position = "bottom",

    legend.text = element_text(size = 8, face = "bold"),
    legend.title = element_text(size = 10, face = "bold"),

    legend.title.align = 0.5
  )
```



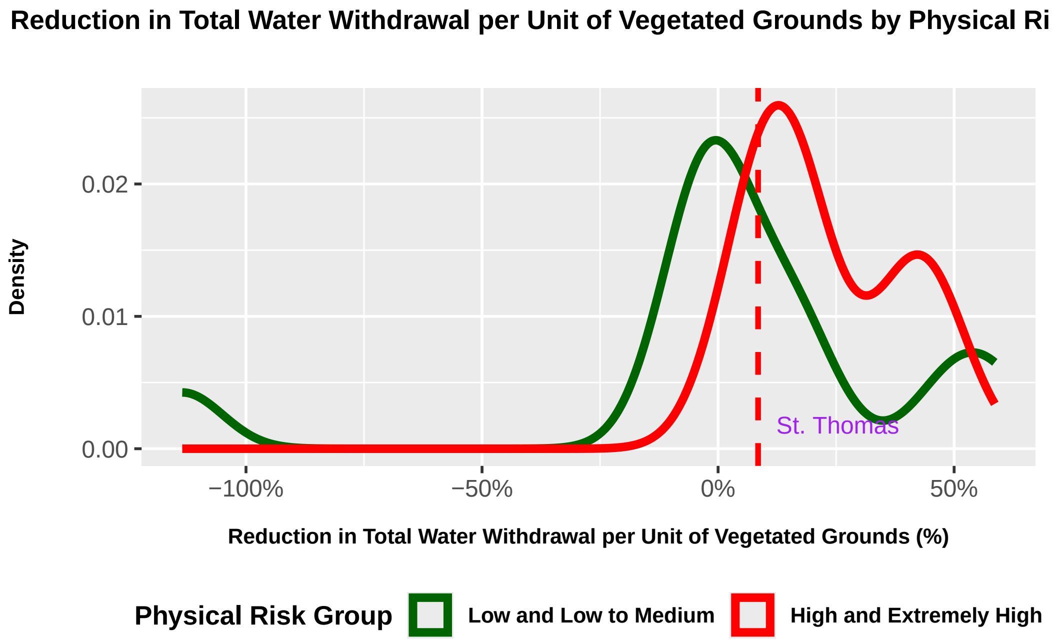
In this data visualization, we are looking at a density chart that displays the frequency of institutions and their percentage reduction in potable water use per weighted campus user. We split up our institutions into two groups, one group of the institutions with the lower half of the physical risk quantity, and the other group is the institutions with the upper half of the physical risk quantity. These two groups are indicated by the red (high risk) and green (low risk) lines along the chart. The the vertical dashed line represents where the University of St. Thomas sits on the distribution for high risk institutions. It is colored red to represent the distribution it corresponds with. In this visualization it is important to note that St. Thomas has actually increased potable water use per weighted campus user since their baseline year. This is something that we could focus on improving in the coming years.

```r
clean2 <- clean2 |>
  mutate(
  percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline = ((potable_wa

ggplot(data = clean2, aes(x = percentage_reduction_in_potable_water_use_per_unit_of_floor_area_
  geom_density(size = 1.5) +
  scale_color_manual(
    values = c('1 (Low and Low to Medium)' = 'darkgreen', '3 (High and Extremely High)' = 'red
    breaks = c('1 (Low and Low to Medium)', '3 (High and Extremely High)'),
    labels = c('Low and Low to Medium', 'High and Extremely High')
  ) +
  labs(
    color = "Physical Risk Group",
    x = "Reduction in Potable Water Use per Unit of Floor Area (%)",
    y = "Density",
    title = "Reduction in Potable Water Use per Unit of Floor Area by Physical Risk Group"
  ) +
  scale_x_continuous(labels = scales::percent) +
  geom_vline(xintercept = 0.2120922, linetype = "dashed", color = "red", size = 1) +
  annotate("text", x = 0.2120922, y = 0.5, label = "St. Thomas",
           color = "purple", vjust = -0.5, hjust = -.15, size = 3) +
  theme(
    plot.title = element_text(size = 10, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),

    legend.position = "bottom",

    legend.text = element_text(size = 8, face = "bold"),
    legend.title = element_text(size = 10, face = "bold"),

    legend.title.align = 0.5
  )
```



In this data visualization, we are looking at a density chart that displays the frequency of institutions and their percentage reduction in potable water use per unit of floor area. We split up our institutions into two groups, one group of the institutions with the lower half of the physical risk quantity, and the other group is the institutions with the upper half of the physical risk quantity. These two groups are indicated by the red (high risk) and green (low risk) lines along the chart. The vertical dashed line represents where the University of St. Thomas sits on the distribution for high risk institutions. It is colored red to represent the distribution it corresponds with. We can see that there are multiple institutions that have increased their water use in this area because they have negative reduction percentages. However, St. Thomas has reduced their water use by 21% which is a substantial decrease. St. Thomas is a bit behind other schools in their group, so this is something that could be improved in the future, but it is still a positive.

```
ggplot(data = clean2, aes(x = percentage_reduction_in_total_water_withdrawal_per_unit_of_vegeta
  geom_density(size = 1.5) +
  scale_color_manual(
    values = c('1 (Low and Low to Medium)' = 'darkgreen', '3 (High and Extremely High)' = 'red
    breaks = c('1 (Low and Low to Medium)', '3 (High and Extremely High)'),
    labels = c('Low and Low to Medium', 'High and Extremely High')
  ) +
  labs(
    color = "Physical Risk Group",
    x = "Reduction in Total Water Withdrawal per Unit of Vegetated Grounds (%)",
    y = "Density",
    title = "Reduction in Total Water Withdrawal per Unit of Vegetated Grounds by Physical Risk
  ) +
  scale_x_continuous(labels = scales::percent_format(scale = 1)) +
  geom_vline(xintercept = 8.473044, linetype = "dashed", color = "red", size = 1) +
  annotate("text", x = 8.473044, y = 0.0005, label = "St. Thomas",
           color = "purple", vjust = -0.5, hjust = -.15, size = 3) +
  theme(
    plot.title = element_text(size = 10, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),

    legend.position = "bottom",

    legend.text = element_text(size = 8, face = "bold"),
    legend.title = element_text(size = 10, face = "bold"),

    legend.title.align = 0.5
  )
```



In this data visualization, we are looking at a density chart that displays the frequency of institutions and their percentage reduction in potable water use per unit of vegetated grounds. We split up our institutions into two groups, one group of the institutions with the lower half of the physical risk quantity, and the other group is the institutions with the upper half of the physical risk quantity.

These two groups are indicated by the red (high risk) and green (low risk) lines along the chart. The the vertical dashed line represents where the University of St. Thomas sits on the distribution for high risk institutions. It is colored red to represent the distribution it corresponds with. We can see that some schools have increased their water use in this area. However, St. Thomas has reduced their water use. This is on par with the majority of the rest of the institutions in their group (high and extremely high risk). St. Thomas has only reduced their water use by 8%, so this is something that could be improved in the future, however it is still a positive.

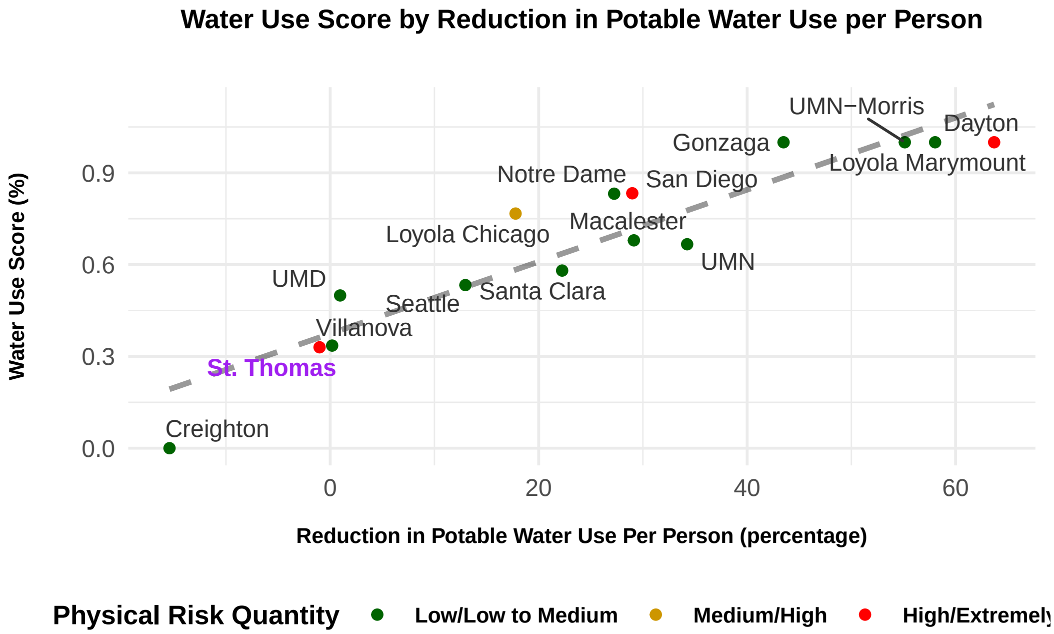
Further exploration of the three percentage variables was warranted, using scatter plots and linear regression to create lines of best fit, understanding St. Thomas' OP-21 performance (in both institution groups) through visual and quantifiable means, respectively. To begin, institution names were shortened to create more space on the charts.

```
clean3 <- clean2 |>
  mutate(
    institution = case_when(
      institution == "Creighton University" ~ "Creighton",
      institution == "Gonzaga University" ~ "Gonzaga",
      institution == "Loyola University Chicago" ~ "Loyola Chicago",
      institution == "Loyola Marymount University" ~ "Loyola Marymount",
      institution == "Santa Clara University" ~ "Santa Clara",
      institution == "Seattle University" ~ "Seattle",
      institution == "University of Dayton" ~ "Dayton",
      institution == "University of Notre Dame" ~ "Notre Dame",
      institution == "University of San Diego" ~ "San Diego",
      institution == "Villanova University" ~ "Villanova",
      institution == "University of St. Thomas" ~ "St. Thomas",
      institution == "Bemidji State University" ~ "Bemidji",
      institution == "Carleton College" ~ "Carleton MN",
      institution == "College of Saint Benedict" ~ "St. Ben's",
      institution == "St. John’s University" ~ "St. John's",
      institution == "Concordia College - Moorhead" ~ "Concordia Moorhead",
      institution == "Macalester College" ~ "Macalester",
      institution == "Winona State University" ~ "Winona",
      institution == "University of Minnesota, Twin Cities" ~ "UMN",
      institution == "University of Minnesota, Morris" ~ "UMN-Morris",
      institution == "University of Minnesota, Duluth" ~ "UMD",
      institution == "Augsburg University" ~ "Augsburg",
      institution == "Concordia in St. Paul" ~ "Concordia St. Paul",
      institution == "Hamline University" ~ "Hamline",
      institution == "St. Kate’s University" ~ "St. Kate's",
      institution == "St. Olaf College" ~ "St. Olaf",
      TRUE ~ institution,
    ),
    percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline = percenta
    percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline = percentage
```

```
library(ggrepel)

lm_model1 <- lm(op_21_percent ~ percentage_reduction_in_potable_water_use_per_weighted_campus_u

coefficients <- coef(lm_model1)
intercept <- coefficients[1]
slope <- coefficients[2]

cat("Equation of the line: y =", round(intercept, 2), "+", round(slope, 2), "* x\n")
```

```
Equation of the line: y = 0.37 + 0.01 * x
```

```
clean3 |>
  ggplot(aes(
    y = op_21_percent,
    x = percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline
  )) +
  geom_point(aes(color = risk_group)) +
  geom_smooth(method = "lm", se = FALSE, color = scales::alpha("black", 0.4), linetype = "dashe
  geom_text_repel(aes(
    label = institution,
    fontface = ifelse(institution == "St. Thomas", "bold", "plain")
  ),
  size = 3,
  color = ifelse(clean3$institution == "St. Thomas", "purple", "gray20")) +
  labs(
    title = "Water Use Score by Reduction in Potable Water Use per Person",
    x = "Reduction in Potable Water Use Per Person (percentage)",
    y = "Water Use Score (%)",
    color = "Physical Risk Quantity"
  ) +
  scale_color_manual(values = c("darkgreen", "#CD9600", "red", "gray20"),
  labels = c("Low/Low to Medium", "Medium/High", "High/Extremely High")) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),
    legend.position = "bottom",

    legend.text = element_text(size = 8, face = "bold"),
    legend.title = element_text(size = 10, face = "bold"),

    legend.title.align = 0.5
  )
```



```
summary(lm_model1)
```

```


Call:
lm(formula = op_21_percent ~ percentage_reduction_in_potable_water_use_per_weighted_campus_use
    data = clean3)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.19269 -0.05625 -0.03230  0.11400  0.18355 

Coefficients:
                                                                                Estimate
```

```
(Intercept)                                                         0.374122
percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline 0.011768
                                                                    Std. Error
(Intercept)                                                           0.043941
percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline   0.001298
                                                                    t value
(Intercept)                                                           8.514
percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline   9.064
                                                                    Pr(>|t|)
(Intercept)                                                         1.12e-06
percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline 5.56e-07

(Intercept)                                                         ***
percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1137 on 13 degrees of freedom
Multiple R-squared:  0.8634,    Adjusted R-squared:  0.8529
F-statistic: 82.15 on 1 and 13 DF,  p-value: 5.56e-07
```

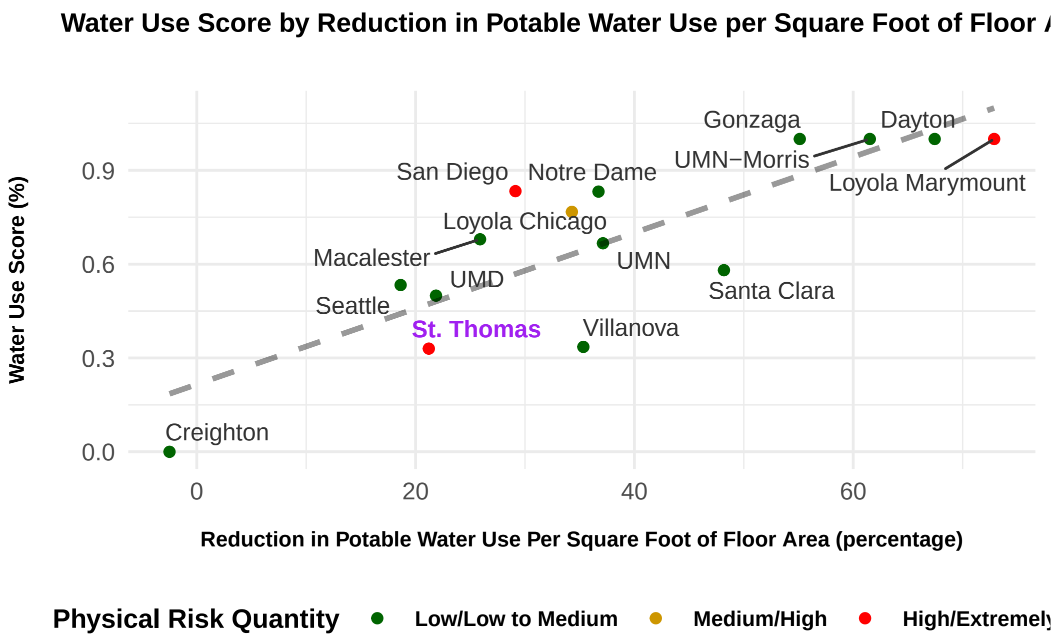
The first scatter plot graphs a school's OP 21 score (as a percent for easier comparison across physical risk quantities) against their percentage reduction in potable water use per person. As seen in the chart, St. Thomas' percentage reduction lags behind most of the schools. The line of best fit has the equation `y= 0.0117x + 0.3741`, indicating that for a 1 unit increase in the percentage reduction in potable water use per person, the model predicts an increase of 0.0117 in the water use score percent (with both coefficients of the equation being statistically significant). The adjusted r-squared value of 0.8529 indicates a strong positive correlation between the x and y variables such that 85.29% of the variability in the water use score (as a percent) is explained by the model using percentage reduction in potable water use per person as a predictor.

This model suggests that St. Thomas stands to gain from investing resources in further reduction of potable water use per person, as other institutions in the two groups were awarded higher scores with greater reduction.

```r
clean3_nooutliers <- clean3


lm_model2 <- lm(op_21_percent ~ percentage_reduction_in_potable_water_use_per_unit_of_floor_are
coefficients <- coef(lm_model2)
intercept <- coefficients[1]
slope <- coefficients[2]

cat("Equation of the line: y =", round(intercept, 2), "+", round(slope, 2), "* x\n")
```

```
Equation of the line: y = 0.22 + 0.01 * x
```

```
clean3_nooutliers |>
ggplot(aes(y=op_21_percent, x=percentage_reduction_in_potable_water_use_per_unit_of_floor_area
  geom_point(aes(color = risk_group)) +
  geom_smooth(method = "lm", se = FALSE, color = scales::alpha("black", 0.4), linetype = "dashe
  geom_text_repel(aes(
    label = institution,
    fontface = ifelse(institution == "St. Thomas", "bold", "plain")
  ),
  size = 3,
  color = ifelse(clean3_nooutliers$institution == "St. Thomas", "purple", "gray20")) +
  labs(
    title = "Water Use Score by Reduction in Potable Water Use per Square Foot of Floor Area",
    x = "Reduction in Potable Water Use Per Square Foot of Floor Area (percentage)",
    y = "Water Use Score (%)",
    color = "Physical Risk Quantity"
  ) +
  scale_color_manual(values = c("darkgreen", "#CD9600", "red", "gray20"),
  labels = c("Low/Low to Medium", "Medium/High", "High/Extremely High")) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),

    legend.position = "bottom",

    legend.text = element_text(size = 8, face = "bold"),
    legend.title = element_text(size = 10, face = "bold"),
    legend.title.align = 0.5
  )
```



```
summary(lm_model2)
```

```


Call:
lm(formula = op_21_percent ~ percentage_reduction_in_potable_water_use_per_unit_of_floor_area_
    data = clean3_nooutliers)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.30847 -0.12095  0.01863  0.12605  0.26464 

Coefficients:
                                                               Estimate
(Intercept)                                                    0.215553
percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline 0.012122
                                                               Std. Error
(Intercept)                                                      0.093976
percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline   0.002217
                                                               t value
```

```
(Intercept)                                                           2.294
percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline   5.468
                                                                    Pr(>|t|)
(Intercept)                                                         0.039113
percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline 0.000108

(Intercept)                                                         *
percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1694 on 13 degrees of freedom
Multiple R-squared:  0.697, Adjusted R-squared:  0.6737
F-statistic:  29.9 on 1 and 13 DF,  p-value: 0.0001078
```

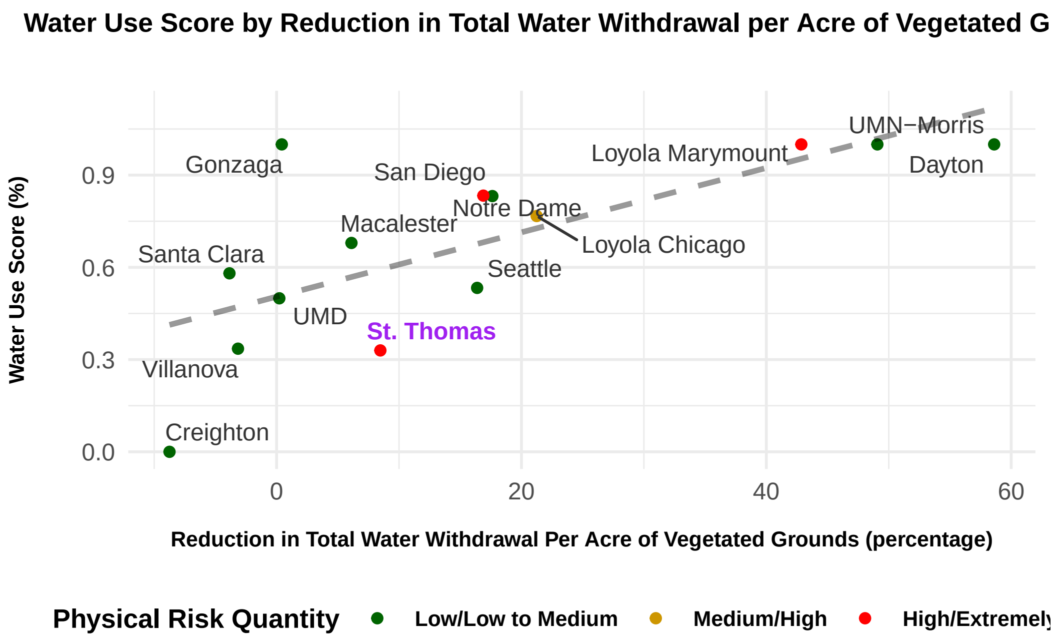
The next scatter plot graphs a school's OP 21 score against their percentage reduction in potable water use per square foot of floor area. Once again, St. Thomas' percentage reduction lags behind most of the schools. The line of best fit has the equation `y= 0.0121x + 0.2155`, indicating that for a 1 unit increase in the percentage reduction in potable water use per square foot of floor area, the model predicts an increase of 0.0123 in the water use score percent (with both coefficients of the equation being statistically significant). The adjusted r-squared value of 0.6737 indicates a strong positive correlation between the x and y variables such that 67.37% of the variability in the water use score (as a percent) is explained by the model using percentage reduction in potable water use per square foot of floor area as a predictor.

This model suggests that St. Thomas stands to gain from investing resources in further reduction of potable water use per square foot of floor area, as other institutions in the two groups were always awarded higher scores with greater reduction (however, it should be noted that schools also received higher scores with less reduction). Although the slope of this equation is slightly greater than the previous equation (0.0121 vs. 0.0117), the smaller adjusted r-squared value (0.6737 vs. 0.8529) suggests that more time invested in improving reduction in water use per square foot of floor area will not yield a greater gain in OP 21 score (as a percent) than time invested in improving reduction in water use per person.

```r
clean3_nooutliers <- clean3 |>
  filter(institution != "UMN")

lm_model3 <- lm(op_21_percent ~ percentage_reduction_in_total_water_withdrawal_per_unit_of_vege
coefficients <- coef(lm_model3)
intercept <- coefficients[1]
slope <- coefficients[2]

cat("Equation of the line: y =", round(intercept, 2), "+", round(slope, 2), "* x\n")
```

```
Equation of the line: y = 0.5 + 0.01 * x
```

```r
clean3_nooutliers |>
ggplot(aes(y=op_21_percent, x=percentage_reduction_in_total_water_withdrawal_per_unit_of_vegeta
  geom_point(aes(color = risk_group)) +
  geom_smooth(method = "lm", se = FALSE, color = scales::alpha("black", 0.4), linetype = "dashe
  geom_text_repel(aes(
    label = institution,
    fontface = ifelse(institution == "St. Thomas", "bold", "plain")
  ),
  size = 3,
  color = ifelse(clean3_nooutliers$institution == "St. Thomas", "purple", "gray20")) +
  labs(
    title = "Water Use Score by Reduction in Total Water Withdrawal per Acre of Vegetated Grou
    x = "Reduction in Total Water Withdrawal Per Acre of Vegetated Grounds (percentage)",
    y = "Water Use Score (%)",
    color = "Physical Risk Quantity"
  ) +
  scale_color_manual(values = c("darkgreen", "#CD9600", "red", "gray20"),
  labels = c("Low/Low to Medium", "Medium/High", "High/Extremely High")) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 10, face = "bold", hjust = 0.5, margin = margin(b = 20)),
    axis.title.y = element_text(size = 8, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 8, face = "bold", margin = margin(t = 10)),

    legend.position = "bottom",

    legend.text = element_text(size = 8, face = "bold"),
    legend.title = element_text(size = 10, face = "bold"),
    legend.title.align = 0.5
  )
```



```
summary(lm_model3)
```

```
Call:
lm(formula = op_21_percent ~ percentage_reduction_in_total_water_withdrawal_per_unit_of_vegeta
    data = clean3_nooutliers)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41291 -0.13164  0.01621  0.11500  0.49104 

Coefficients:
                                                                            Est
(Intercept)                                                                 0.50
percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baseline 0.01
                                                                            Std
(Intercept)                                                                   0
percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baseline   0
                                                                            t va
```

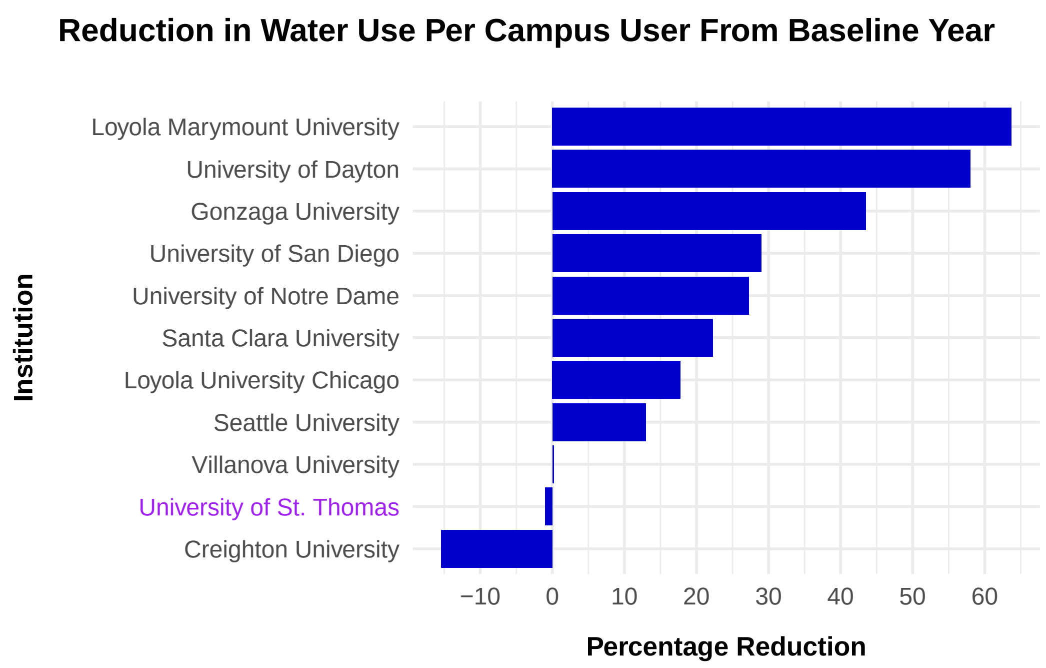
```
(Intercept)                                                                            6
percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baseline   3
                                                                                     Pr(
(Intercept)                                                                          2.5
percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baseline  0.0

(Intercept)                                                                          ***
percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baseline **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.225 on 12 degrees of freedom
Multiple R-squared:  0.5063,    Adjusted R-squared:  0.4652
F-statistic: 12.31 on 1 and 12 DF,  p-value: 0.004317
```

The last scatter plot graphs a school's OP 21 score against their percentage reduction in potable water use per acre of vegetated grounds. The -113 percentage reduction of UMN Twin Cities was deemed and outlier and removed from the data set. St. Thomas' percentage reduction lags behind some of the schools, but not as many as the previous graphs. The line of best fit has the equation `y= 0.0105x + 0.5045`, indicating that for a 1 unit increase in the percentage reduction in potable water use per acre of vegetated grounds, the model predicts an increase of 0.0105 in the water use score percent (with both coefficients of the equation being statistically significant). The adjusted r-squared value of 0.4652 indicates a strong positive correlation between the x and y variables such that 46.52% of the variability in the water use score (as a percent) is explained by the model using percentage reduction in potable water use per acre of vegetated grounds as a predictor.

This model suggests that St. Thomas stands to gain from investing resources in further reduction of potable water use per square foot of floor area, as other institutions in the two groups were always awarded higher scores with greater reduction (however, it should be noted that schools also received higher scores with less reduction). The slightly smaller slope (0.0105 vs. 0.0123 vs. 0.0117) and much lower adjusted r-squared value (0.4652 vs. 0.6737 vs. 0.8529) of this equation than the previous equations suggests that more time invested in improving reduction in water use per acre of vegetated grounds will not yield a greater gain in OP 21 score (as a percent) than the other percentage reduction variables.

The goal of the final set of visualizations is to see exactly where the University of St. Thomas lines up among the Catholic Benchmark Institutions for the three variables we chose to analyze. These three bar charts show the percent change of water use arranged in numeric order with St. Thomas highlighted in purple, making it easy to see where St. Thomas stands among its peers.

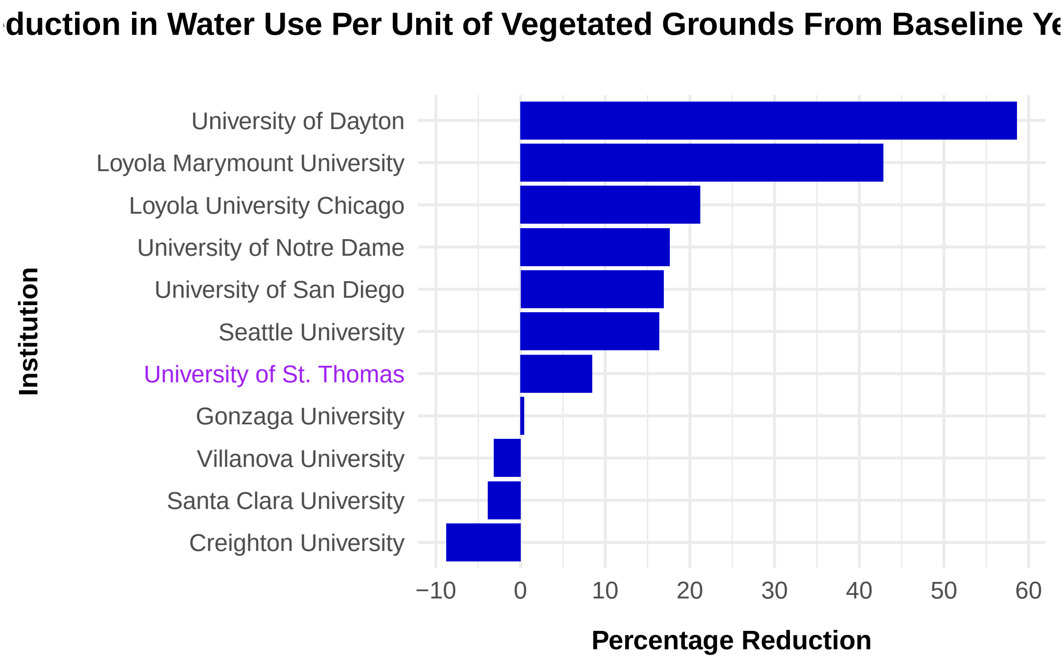
```
clean3 <- clean2 |>
  filter(institution %in% catholic_benchmark_institutions) |>
  mutate(
  percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline = percentag
    label_color = ifelse(institution == "University of St. Thomas", "purple", "gray30")
  )

ggplot(clean3, aes(
  y = reorder(institution, percentage_reduction_in_potable_water_use_per_weighted_campus_user_
  x = percentage_reduction_in_potable_water_use_per_weighted_campus_user_from_baseline,
)) +
  geom_bar(stat = "identity", fill = "blue3") +
  labs(y = "Institution", x = "Percentage Reduction", title = "Reduction in Water Use Per Campu
  scale_x_continuous(breaks = seq(-20, 100, by = 10)) +
  theme_minimal() +
  theme(
        plot.title = element_text(size = 12, face = "bold", hjust = 0.5, margin = margin(b = 2
    plot.title.position = "plot",
    axis.title.y = element_text(size = 10, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 10, face = "bold", margin = margin(t = 10)),

    axis.text.y = element_text(color = clean3 |> arrange(percentage_reduction_in_potable_water_
  )
```



This bar chart shows the percent change in water use per weighted campus user from the baseline year to the performance year for schools in the Catholic Benchmark Institutions group. Most schools in this group were able to decrease their water use per weighted campus user with some schools cutting their water use by more than half. Only two schools actually increased their water use, having a negative percentage reduction in water use. One of these two schools is the University of St.Thomas. They had the second worst change in water use per campus users out of all other schools in this group. Their change in water use in this category was an increase of about 1% from the baseline year to the performance year.

Although a 1% change is not a dramatic increase, it is certainly the wrong direction to go for a school that wants to become sustainable. The increase in water use per campus user is the part of the reason why the St. Thomas received such a low OP-21 water use score and was put into the "High" Physical Risk Quantity group. If St. Thomas wants to become a sustainable institution and receive a higher AASHE STARS rating they will need to put more effort towards decreasing water use per campus user and increasing their water use score.

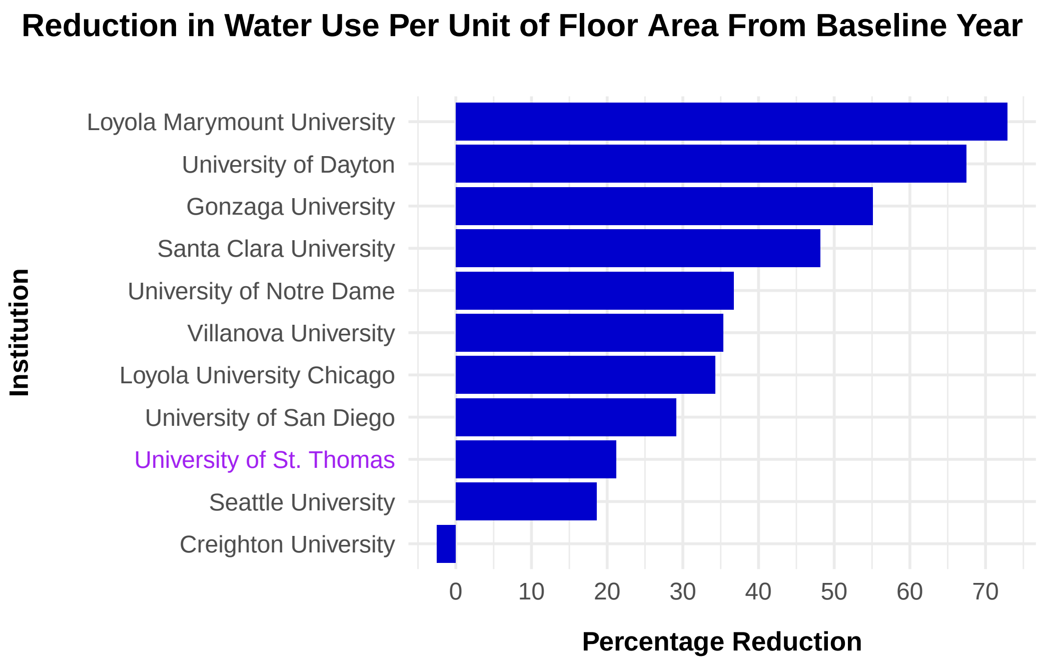
```
clean3 <- clean2 |>
  filter(institution %in% catholic_benchmark_institutions) |>
  mutate(
    label_color = ifelse(institution == "University of St. Thomas", "purple", "gray30")
  )

ggplot(clean3, aes(
  y = reorder(institution, percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated
  x = percentage_reduction_in_total_water_withdrawal_per_unit_of_vegetated_grounds_from_baselin
)) +
  geom_bar(stat = "identity", fill = "blue3") +
  labs(y = "Institution", x = "Percentage Reduction", title = "Reduction in Water Use Per Unit
  scale_x_continuous(breaks = seq(-20, 100, by = 10)) +
  theme_minimal() +
  theme(
        plot.title = element_text(size = 12, face = "bold", hjust = 0.5, margin = margin(b = 20
    plot.title.position = "plot",
    axis.title.y = element_text(size = 10, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 10, face = "bold", margin = margin(t = 10)),

    axis.text.y = element_text(color = clean3 |> arrange(percentage_reduction_in_total_water_wi
  )
```



This bar chart shows the percent change in water use per unit of vegetated grounds from baseline year to performance year. Most institutions from the Catholic Benchmark Institutions group were able to decrease their water use per area of vegetated grounds, though a few had an increase in this category. The University of St. Thomas was able to decrease water use for this metric, but of the institutions that made a decrease, St. Thomas had the smallest change. The seven other institutions that had a decrease were able to reduce water use per unit of vegetated grounds by more than 15% with one institution decreasing by almost 60%. St.Thomas, on the other hand, produced a decrease of only 8.5%.

Although the University of St. Thomas was able to decrease water use per unit of vegetated grounds from the baseline year to the performance year, the decrease was not as large as most other schools from the Catholic Benchmark Institutions. This is a category that St. Thomas will surely need to put more focus on if they wish to receive a higher water use score.

```r
clean3 <- clean2 |>
  filter(institution %in% catholic_benchmark_institutions) |>
  mutate(
  percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline = percentage_i
    label_color = ifelse(institution == "University of St. Thomas", "purple", "gray30")
  )

ggplot(clean3, aes(
  y = reorder(institution, percentage_reduction_in_potable_water_use_per_unit_of_floor_area_fro
  x = percentage_reduction_in_potable_water_use_per_unit_of_floor_area_from_baseline,
)) +
  geom_bar(stat = "identity", fill = "blue3") +
  labs(y = "Institution", x = "Percentage Reduction", title = "Reduction in Water Use Per Unit
  scale_x_continuous(breaks = seq(-20, 100, by = 10)) +
  theme_minimal() +
  theme(
        plot.title = element_text(size = 12, face = "bold", hjust = 0.5, margin = margin(b = 20
    plot.title.position = "plot",
    axis.title.y = element_text(size = 10, face = "bold", margin = margin(r = 20)),
    axis.title.x = element_text(size = 10, face = "bold", margin = margin(t = 10)),

    axis.text.y = element_text(color = clean3 |> arrange(percentage_reduction_in_potable_water_
  )
```


Reduction in Water Use Per Unit of Floor Area From Baseline Year

The final bar chart shows the change in water use per unit of floor from the baseline year to the performance year for institutions in the Catholic Benchmark Institutions group. Compared to the ten other institutions here, the University of St. Thomas has the third poorest percentage change for this category. St. Thomas still had a much larger percentage decrease in this category than it had in other categories with a 21.2% reduction in water use per unit of floor area, but the mean percentage reduction in this category for Catholic Benchmark Institutions was substantially higher at 37.9%.

The University of St. Thomas was able to decrease their water use per unit of floor area by a fair amount, but they are still falling behind other Catholic Benchmark Institutions in this category. To receive a higher water use score in the future and become a more sustainable institution, St. Thomas should seek to reduce their water use per unit of floor area.

The analysis of the AASHE STARS water use data has allowed us to identify, compare, and predict the OP-21 credit score for St. Thomas and other benchmark institutions of interest. By studying the three underlying parts that comprise a school's OP 21 score, a better understanding of how an institution such as St. Thomas can improve their performance was gained. Further examination of three important variables regarding reduction in water use was important for discovering why St. Thomas's OP-21 water use score is so low. Several data visualizations were created from the variables "reduction in water use per weighted campus user from baseline", "reduction in water use per unit of vegetated ground from baseline", and "reduction in water use per unit of floor area from baseline". These visualizations effectively illustrated that St. Thomas consistently lags behind their peers in these three categories. For change in water use compared to the number of weighted campus users, St. Thomas actually saw an increase in water use from the baseline year. The creation of simple linear regression models for the three variables against OP-21 score allowed for understanding of which variables had higher predictive power for an institutions OP-21 score. All models had close to the same slope, but the model using water use per weighted campus user had by far the highest adjusted R-squared value, indicating that St. Thomas should put the greatest amount of focus in this category. To receive a higher OP-21 water use score, a higher AASHE STARS rating, and ultimately become a more sustainable institution, St. Thomas should seek to reduce their water use per unit of floor area, their water use per unit of vegetated grounds, and most importantly their water use per number of weighted campus users.