SYE Report

Trent Meyer

2023-05-08

Contents

SYE Introduction

When considering what to do for my SYE, I had two options. The first being to complete the rest of my SLU Fellowship research project: “Assessment of Water Bottle Filling Station Use on St. Lawrence University Campus”, which I had started back in Spring 2021. After presenting at the Laurentian Weekend poster session, I took a look to see what I had left, and there was not a lot to do. With so little left to do, I began looking for smaller research projects that I could complete by the end of the semester. With the 2022 World Cup around the corner, I began to see many predictions and conversations about who was going to win. I wanted to see how the predictions worked, and this brought me to the second part of my

SYE. I found an article from FiveThirtyEight that went over the basics of their predictions, and I wanted to recreate that using the data from their GitHub. This took much longer than expected, as I had a lot to learn about iterations, loops, and creating your own functions in R. By the end, I was able to recreate, to an extent, their table of probabilities that each team in the World Cup would advance out of the group states of the tournament.

Assessment of Water Bottle Filling Station Use on St. Lawrence University Campus

Introduction

With the ever-growing population, water demand will increase by almost 40% by the year 2050, and the ability to access clean, safe drinking water will become even more of an issue (Tian 2019). One popular source of drinkable water is disposable, plastic water bottles. However, there are a multitude of issues associated with these plastic water bottles. Some estimates say that these will take centuries to decompose (Steinmetz 2013). Steinmetz (2013) also mentioned that plastic waste accumulated to 32 million tons produced in just the year 2011, and that in the city of San Francisco, only 13% of plastic was actually being recycled. In one year, there are approximately 50 billion plastic water bottles being used (Lappé 2016).

As a result of the problems associated with single use plastic water bottles, many communities and schools began looking for ways to reduce plastic water bottle usage, while also providing clean water that is safe to drink. One of the solutions that has become popular in many recreation centers and school campuses is water bottle filling stations, sometimes called "hydration stations." Many of the hydration stations use carbon filters to improve the taste and filter the tap water that the spout dispenses (Brandon 2011). Using water bottle filling stations, the water bottle can be filled directly to the top, unlike water fountains where you have to tip your bottle to fill it (Marohn 2011). Marohn (2011) describes that the stations allow you to easily refill water bottles, which eliminates the need to constantly buy a new water bottle. The water bottle filling stations are tall enough so that users can set a bottle underneath of the spout, meaning it can get filled all the way to the top (Heldt 2012). The water bottle filling stations have given us the opportunity to fill and reuse plastic water bottles many times, meaning we are able to avoid using an unnecessary number of disposable plastic water bottles. Heldt (2012) explained that many of the water bottle filling stations have counters to track how many plastic water bottles are being avoided by refilling a reusable bottle with the station.

A common issue is that bacteria can grow on many surfaces we would not normally expect. An example of bacteria that can grow in water sources is the Legionella bacteria, which causes the disease known as Legionnaires' disease (Sandeep 2018). Sandeep (2018) described the origin of the name "Legionnaires" which was linked to an American Legion convention in 1976. The Legionella bacteria grew on the water source, which ultimately affected a total of 182 individuals, and killed 29 of those (Sandeep 2018). This is just one of the many issues with water sources, as many are often said to be clean and filtered, but are they as clean as we truly think they are? This question was being investigated by Dr. Lorraine Olendzenski at St. Lawrence University in the Spring of 2021. She, along with a group of five students, sampled water from the water bottle filling stations on campus. They discovered multiple species of bacteria were present on the spout and base of the filling stations.

We chose to parallel Dr. Olendzenski's work with two surveys to investigate the usage of water bottle filling stations on campus by St. Lawrence University students. We administered two online surveys that were offered to St. Lawrence students over the age of 18, both those who are on-campus, those who were remote for the Spring 2021 semester. The surveys aimed to determine the frequency of use and students' perception of the quality and safety of the water bottle filling stations. We asked questions about students' general health including sleeping and dietary habits, along with their stress levels while on campus. Further, we wanted to determine if more students that use the water bottle filling stations undergo more specific health issues. However, one difficulty of studying disease spread is that there are a multitude of reasons someone could get sick, meaning disease is multi-factorial. This led a shift in our motivation for doing the follow-up survey, instead focusing on accessibility to and perception of the water bottle filling stations on campus. This approach was more asking why someone was using the filling station, or if they could not, why they were unable to do so.

Methods

Design

The online surveys, which were administered through Qualtrics, included 25 questions in 2021, and 20 questions in 2022 (questions available on on the project's GitHub Repository) and was only offered to St. Lawrence University students over the age of 18. The surveys were open for one week in March 2021 and April 2022, respectively. With some students remotely taking classes during the Spring 2021 semester, the survey was also offered to them to maintain proper representation of the student body on campus. The survey proposal was submitted to and accepted by the IRB on February 25, 2021. Before taking both surveys, participants had to read and accept the informed consent statement (also available on the project's GitHub Repository). Responses were analyzed and presented in aggregate (not individually). Upon completion of the surveys, participants were eligible for a drawing for one of three Amazon gift cards ($50 each in 2021, $25 each in 2022). If participants chose to participate in the drawings, a link at the end of the surveys took them to a separate Qualtrics site, where they were asked to provide their name and e-mail address for delivery of the gift card. This information was NOT linked to the participants' individual survey answers. Three respondents were randomly selected for each survey and received their gift card by e-mail within three weeks of the survey closure.

Implementation

Fliers, which included a brief description of the survey and a QR code, were posted in frequently traveled areas and residential buildings on campus. These frequently traveled areas included, but were not limited to, Dana Dining Hall, Sullivan Student Center, and ODY Library. During the Spring 2021 survey, a digital flier was emailed to the three Residential Coordinators which they distributed to their own Community Assistants. Community Assistants were also given the opportunity to post these fliers on their individual floors in each residential hall. Stacey Olney LaPierre, the Senior Associate Director of Residence Life and Housing Operations, was able to contact those who were living off campus via emailing a digital copy of the flier along with a link. Sharon Rodriguez, the Residential Coordinator in charge of Greek and Theme houses, sent an email to the students that she oversees. This provided these upperclassmen with the opportunity to participate. Allowing those off-campus to participate was part of an effort to maintain proper representation of the whole student body, rather than just those living on-campus in dorms.

The digital flier was also sent to the Class Instagram pages. The class Instagram pages were able to either post the flier directly to their profile or include it on their profile story. The digital flier was also sent to the St Lawrence Instagram to be featured on the "This Week in Posters" story along with a link to the survey.

Only the student researcher and principal investigator had access to the recorded data, which was stored and secured in the Qualtrics software program with password protection. The recorded data was exported to a password protected Google Drive for further statistical analysis, again only accessible by the student researcher and principal investigator. The collected data was only included in internal St. Lawrence University poster sessions, along with presentations for various clubs on campus.

Statistical Analysis

The raw data was downloaded from the Qualtrics software into an Excel spreadsheet. In order to run statistical tests, the data in a ".CSV" file was uploaded from an Excel spreadsheet to R. In R, we were able to create either bar and lollipop plots or tables for each question. For associations between filling

Table 1: Gender Demographics

gender	prop2021	prop2022
Female	0.77	0.66
Male	0.22	0.30
GQGNC	0.01	0.04

Table 2: Class Year Demographics

classyear	prop2021	prop2022
First-Year	0.28	0.41
Sophomore	0.27	0.31
Junior	0.23	0.15
Senior	0.22	0.11

frequency and specific illnesses, we used a linear-by-linear association test. This association test ensured that the categorical predictors were seen as ordinal and not nominal. We got a p-value from this association test, and then used the dichotomous method where we determined if the association was significant based on whether or not the p-value was greater than 0.05.

Results

Firstly, the habits surrounding drinking water and water bottles in general were examined. The survey results reported that in both 2021 and 2022 approximately 89% of survey respondents consumed 1 Liter of water or more per day (Fig. 1). Relating to frequency of filling station usage, in 2021 just over half of the survey respondents reported using the water bottle filling station at least once each day, compared to 83.2% in 2022! (Fig. 2).

Respondents were asked to report their level of agreement with six statements relating to the water bottle filling stations (Fig. 3). “Somewhat agree” and “Strongly agree” responses were grouped into “Agree”; while the same was done for the “Disagree” responses (Fig. 3).

First, they were asked to report their level of agreement with whether the water from the filling stations was either more chemically pure, safer to drink, along with if the filling stations were clean (Fig. 3). Next, the respondents were asked about specific functions of the water bottle filling stations. They were asked to report their agreement with whether the water bottle filling stations prevent the usage of disposable plastic bottles, made it easier to drink water, and provided better tasting water (Fig. 3).

Figure 4 shows the most frequently used locations on campus in each year. For both years, the top two were the location on the first floor of the student center and the first floor of the library, referred to as “ODY.” The location outside of Newell Fieldhouse and the fitness center was the third most frequently used in 2021, and fourth most in 2022. We see the location on the third floor of the student center, put in before the Fall 2021-Spring 2022 academic year, was quite popular (Fig. 4).

Respondents were asked for their additional thoughts regarding the filling stations, and their responses were analyzed for key words. The most common key words in 2021 were “more”, “convenience”, and “filter.” In 2022, we still see “more” as the most common, however “filter” and “convenience” were the next most common.

Discussion

One thing to note is that the motivation behind the surveys shifted away from determining if illness was caused by filling station usage. It was almost impossible to pinpoint exactly which filling station could

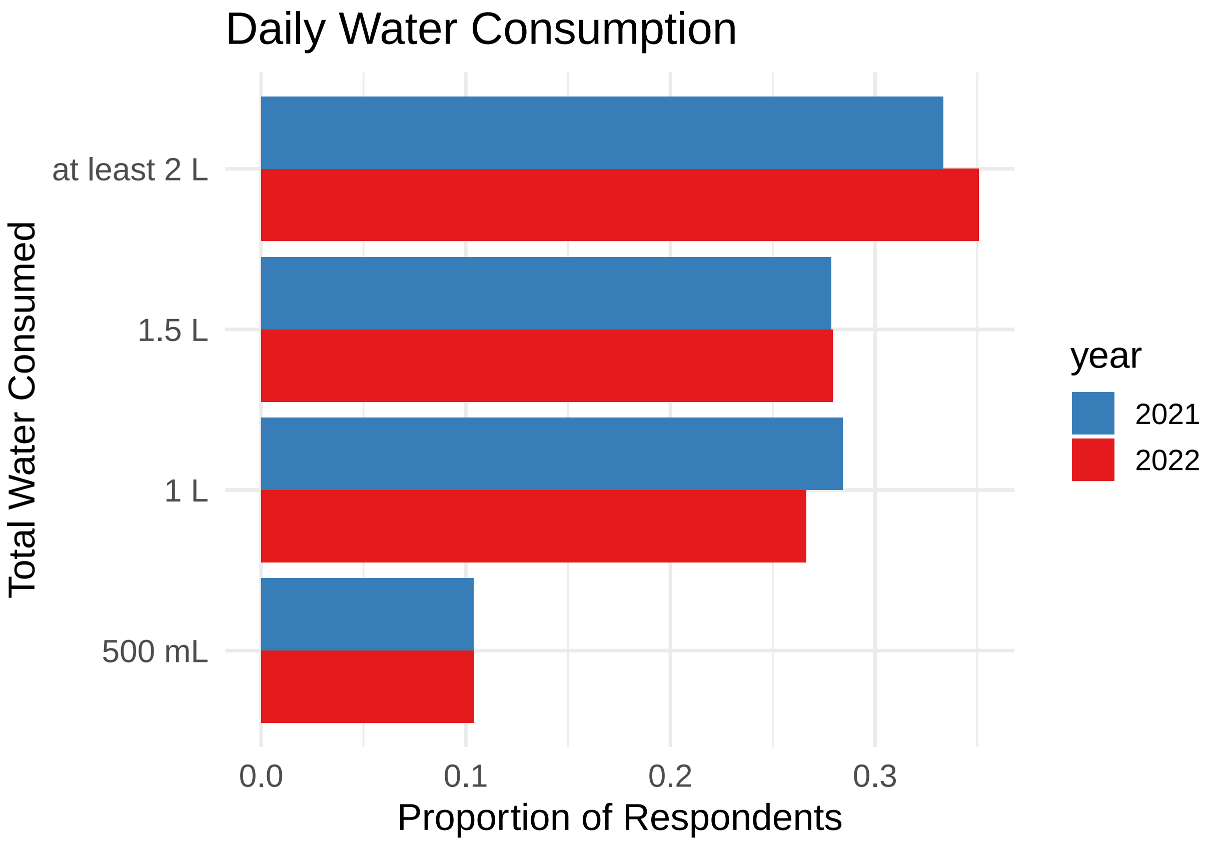

Figure 1: Self-reported daily water consumption

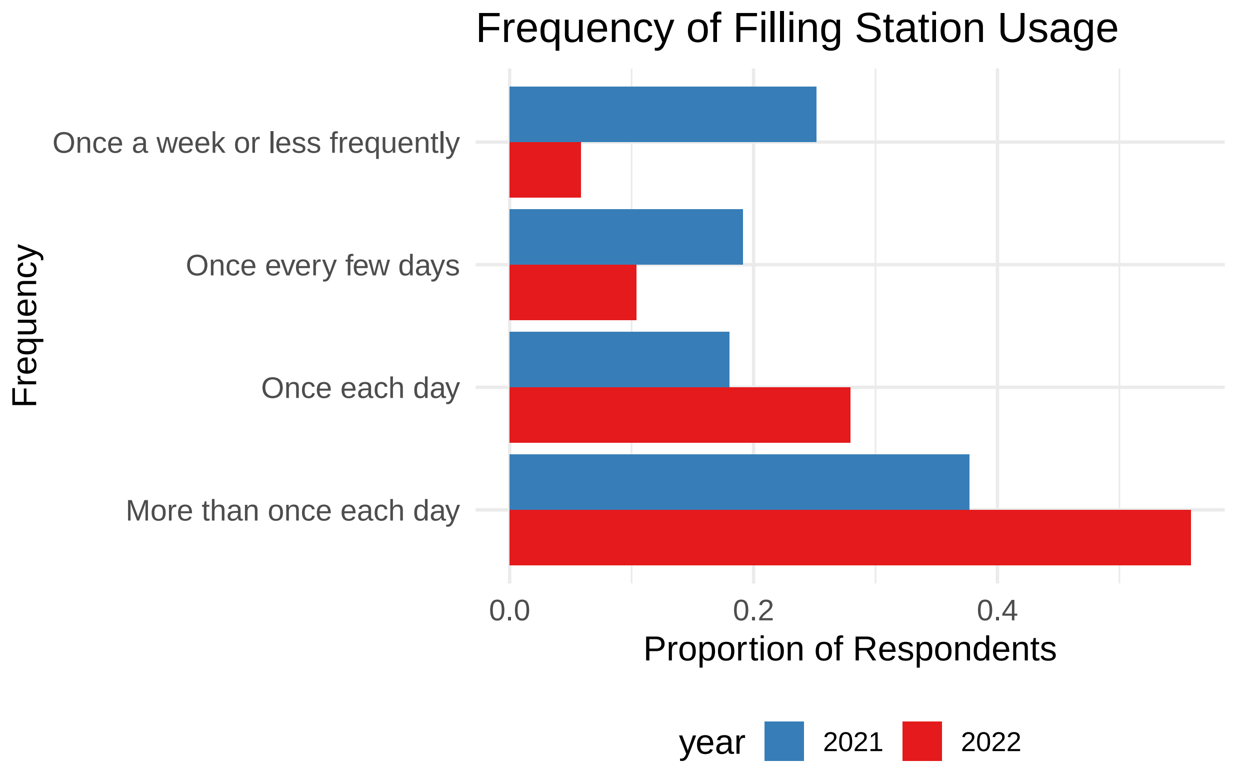

Figure 2: Self-reported filling station frequency of use (2021 n = 186, 2022 n = 155)

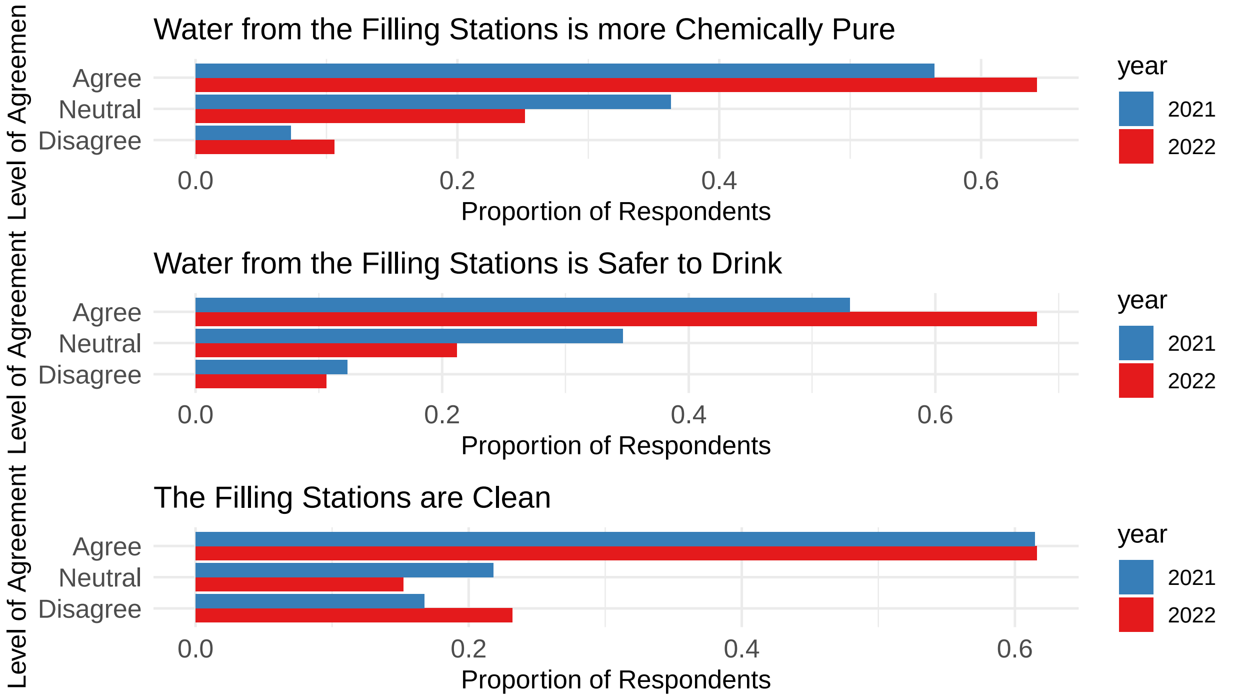

Figure 3: Respondents' perception of filling stations, pooled "Somewhat agree" and "Strongly Agree" responses as "Agree"; the same was done for the "Disagree" responses

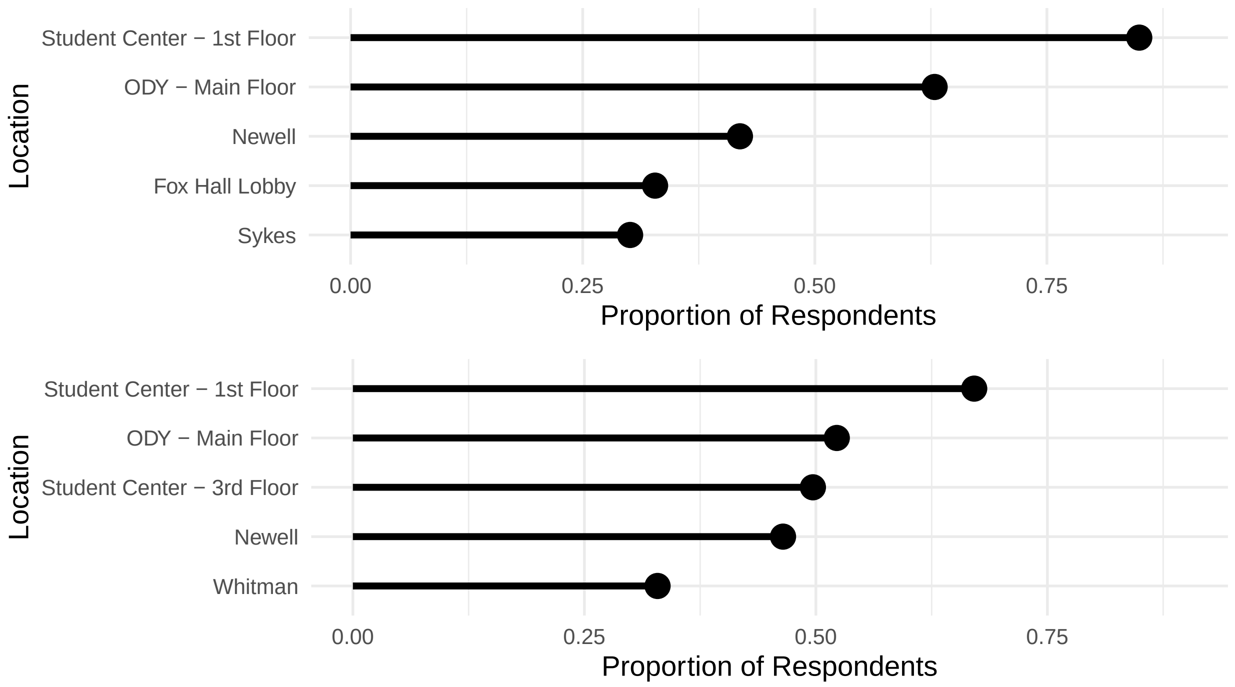

Figure 4: Top five most used filling stations each year, top: 2021, bottom: 2022

have caused someone to become ill. Also, we can not be completely sure that the filling station even caused illness, as illnesses are multi-factorial. Our motivation shifted more towards accessibility and perception of filling stations on campus, rather than the illness question.

One obvious thing we can understand is that students at St. Lawrence University use the filling stations. They care about where their water comes from, and expressed lots of interest in these surveys. They seemed to show lots of concern about the cleanliness and maintenance of the filling stations, as well as the accessibility across campus. One issue is that not all residential buildings has these filling stations. During the Fall 2020-Spring 2021 semesters, students were mostly not allowed to access other residential buildings besides the one they lived in. This created issues for students who did not have a filling station in their building.

Students were also concerned about accessibility throughout the day where they have their classes. Some students who have classes in say Atwood Hall have access, while others in Johnson Hall of Science do not, meaning they have to walk to other buildings just to use a filling station. Accessibility to water throughout the day is important to all students, specifically athletes, who drink lots of water. Students here try their best to avoid plastic, single-use water bottles, so one can see almost everyone carrying around a reusable bottle. The older regular water fountains do not fill up bottles all the way, and many students simply avoid them because they are outdated.

Adding more filling stations across campus is something that many survey respondents agreed on. However, they also cared about the maintenance and cleanliness of them, meaning their filters need to be replaced more frequently, and maintenance be completed in a timely manner, rather than weeks after. Accessibility to water is something many take for granted, and St. Lawrence University was shown how difficult it was to access free, clean water during the COVID-19 pandemic.

World Cup Prediction

Introduction

The FIFA World Cup is an event that grips the attention of soccer (or fútbol, football depending on where you are in the world) fans all across the world. This event occurs once every four years, bringing together millions of individuals from a wide range of cultures all over the world. To have your nation qualify for the World Cup is quite prestigious as only 32 teams (increased to 48 teams in 2026) are able to do so. This rigorous qualifying schedule involves playing the top teams in your region over the course of about two years prior to the next World Cup.

The United States, despite not being completely enthralled like say Brazil, still nonetheless sees a magic atmosphere surrounding the tournament. However, in 2018 the United States lost to Trinidad & Tobago, a team they were supposed to easily beat, in their last qualifying game which effectively knocked them out of that cycle's tournament. This sparked quite a long chain of events in the United States Soccer Federation (USSF, the main governing body for soccer and the national teams) over the next four years. This included completely revamping the youth development system, hiring a new manager, and phasing out all of the aging players in favor for bright, up and coming stars making their presence known in European academies.

This chain of events changed quite a lot for the USSF, as seen since 2018 and today. Our men's national team saw its stars playing at some of the most highly regarded clubs in the world such as Juventus, Chelsea, and Borussia Dortmund. Not only to mention the quality of the clubs, these players were teenagers starting and making an impact for their club. We saw a higher number of teenagers playing in Europe than ever before, meaning our players were developing at the best academies in Europe. Over the next few years, our young team managed to win two regional tournaments and secure our spot in the 2022 World Cup in Qatar.

One thing that has gained popularity over the past few years is increasing data analytics of large-scale sporting events such as the World Cup. With such a large event, many want to predict the winner of this prestigious trophy. After 22 tournaments, only eight have managed to win the tournament. Many companies such as FiveThirtyEight and ESPN, to name only a couple, attempt to predict the winner of the upcoming tournament using many different methods.

FiveThirtyEight created a process to predict the winner of the 2018 tournament. This process is explained, along with the data being available in a GitHub repository. Using their data, I am attempting to recreate the probabilities that each team will win each game in the group stage, and use this to determine who is most likely to make it past the group stage. I will also use this same process for the data from the 2022 tournament, which has currently been ongoing in the months of November and December.

Methods

To begin their predictions, each team's World Cup roster was given an SPI rating, which judged the strength of that team. Each team was given an offensive rating, which is "the number of goals that it would be expected to score against an average team on a neutral field" (Boice, 2018). The defensive rating followed the same situation, except the number of goals they would be expected to concede against an average team. The overall "match" SPI rating is the "percentage of points . . . the team would be expected to take if the match were played over and over again" (Boice, 2018). This SPI rating is calculated using a database of international matches dating back to 1905! They also calculated a roster-based SPI rating which looked at each player, and how much they played, in what league, and how elite the competition they are facing each week is.

Once the SPI rating was calculated, they began by predicting the match scores, which are the number of goals that team would need to score to uphold their rating. Using the projected score, they completed a Poisson process where they predicted the probabilities that each team would score zero goals, one goal, two goals, and so on up to ten goals (a little uncommon, but not outlandish).

The next step was to convert these to matrices, and multiply them together to determine the likelihood that one team would score x amount of goals *and* the other team would score y amount of goals. Using this matrix, I was able to calculate the cumulative probability, and generate a random value that would correspond to an outcome based on the matrix. Each match in the group stages was iterated 1000 times, from which we could determine the number of points each team would score in that iteration. Next, I was able to calculate the proportion of times that each team would advance out of the group stages (place in the top two out of four teams in their group).

Results

2018 World Cup

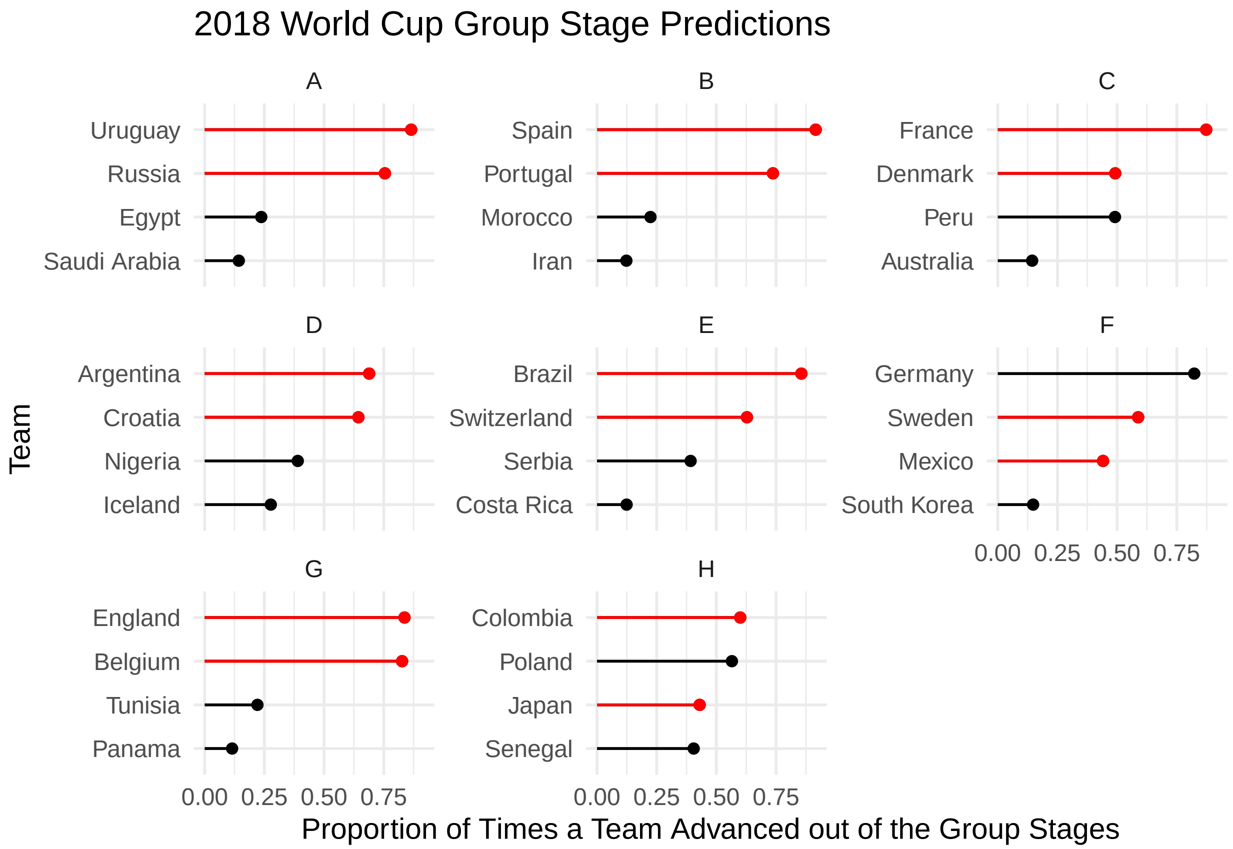

Figure 5: Proportion of times each team advanced out of their respective group

From Figure 5, we are able to see that the teams that made it through the highest proportion of times were Spain (0.916), France (0.873), and Uruguay (0.865). Possibly the most surprising piece of information however is that Germany were the favorites for their groups, however did not actually make it out of the group stage, with Sweden and Mexico advancing instead. The only teams that were expected to advance and did not were Poland and Germany. France were one of the favorites, and end up beating Croatia in the final and winning the whole tournament.

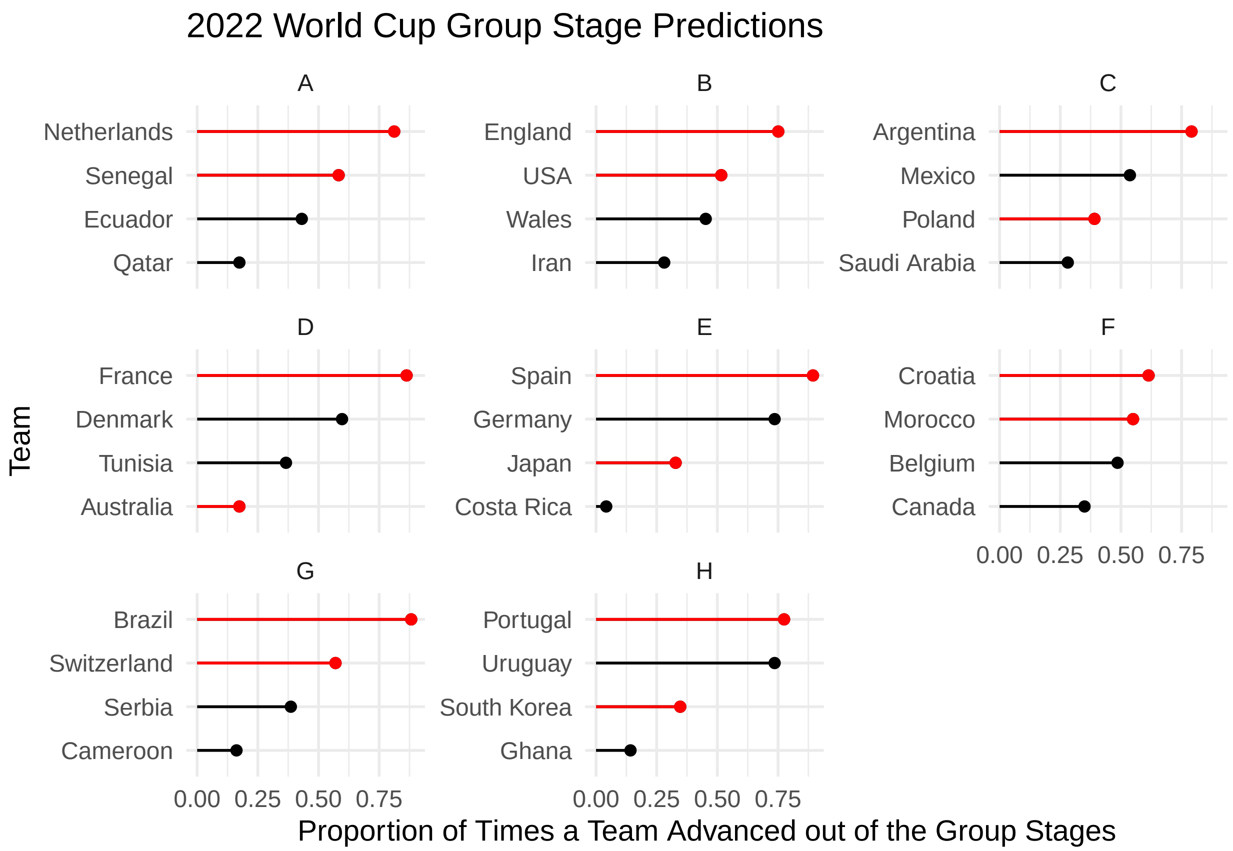

Figure 6: Proportion of times each team advanced out of their respective group

2022 World Cup

From Figure 6, we are able to see that the teams that made it through the highest proportion of times in 2022 were Spain (0.894), Brazil (0.882), and France (0.863) Possibly the most surprising piece of information however is that Germany again were the favorites for their group, however did not actually make it out of the group stage, with Spain and Japan advancing instead. This year had quite a lot more upsets, with four teams who were expected to advance not doing so: Mexico, Denmark, Germany, Uruguay. Australia who were predicted last in the group surprisingly made it out of their group.

Discussion

One of the more difficult aspects of this was that I was not able to completely replicate the probability matrix, as their article only gave limited information on the process. However, with the information given, I learned a lot about predicting World Cup games using skills I learned in linear algebra, probability, and data science. I had to begin by predicting only one game, and once that process was completed, I was able to expand out to one group of four teams. Then once one group was done, I was able to expand to the whole group stage, with eight groups.

The 2022 tournament is at a slightly different time compared to the 2018 version, as the COVID-19 pandemic and the heat in Qatar pushed the tournament into November, rather than June and July. Many players are midway through their seasons, which caused quite a few injuries to star players. We have seen the 2022 tournament be one of the most exciting to watch because of players such as Lionel Messi, Cristiano Ronaldo, and Neymar Jr playing in their last World Cup. Also, this tournament's upsets have shocked many, with some favorites being knocked out in the group stages.

This process took much longer than I expected, and given I only had a semester to work, I was only able to predict the group stages. If given more time, I would be able to figure out a way to set up the knockout stage bracket, and then use the expected scores given to predict who would win that World Cup. This would be quite difficult as you would need to find a way to include the tie breakers, which include point, goal differential, and yellow card accumulation. Once this was completed, you would need to set up a bracket, where the first placed team in one group plays the second placed team in another group.

Code Appendix

```
knitr::opts_chunk$set(echo = TRUE)
library(here)
library(kableExtra)
library(tidyverse)
load(here("data/survey_full.rda"))
summarytable <- function(inputvariable) {
  group_var <- enquo(inputvariable) ## create a "quosure" that puts the
  ## variable you want to group_by in quotes
  ## in the next line, !! then unquotes the variable again for use in
  ## the group_by() function
  survey_full %>% group_by(!!group_var, year) %>% summarise(totalrespondents = n()) %>%
    mutate(yeartotal = if_else(year == 2021,
                               true = 186,
                               false = 155)) %>%
    mutate(proportion = totalrespondents/yeartotal)}

survey_full <- survey_full %>%
  mutate(gender = fct_recode(gender, GQGNC = "Genderqueer / Gender Non-Conforming"))
## recode this level to GQGNC to save space on the table

demtable22 <- survey_full %>%
  filter(year == 2022) %>%
  group_by(gender, year) %>%
  summarise(totalrespondents = n()) %>%
  mutate(yeartotal = if_else(year == 2021,
                              true = 186,
                              false = 155)) %>%
  mutate(prop2022 = totalrespondents/yeartotal) %>% select(1, 5)
## create the table for proportion of each
## level of 2022 respondents for the gender question

survey_full %>% filter(year == 2021) %>%
  group_by(gender, year) %>%
  summarise(totalrespondents = n()) %>%
  mutate(yeartotal = if_else(year == 2021,
                              true = 186,
                              false = 155)) %>%
  filter(gender != "Prefer not to state") %>%
  mutate(prop2021 = totalrespondents/yeartotal) %>%
  select(1, 5) %>% ungroup() %>% mutate(prop2022 = demtable22$prop2022) %>%
  kable(caption = "Gender Demographics", digits = 2)
## create the column for proportion of each
## level of 2021 respondents for the gender question
## and add that to the 2022 proportions


classtable22 <- survey_full %>% filter(year == 2022) %>%
  group_by(classyear, year) %>%
  summarise(totalrespondents = n()) %>%
  mutate(yeartotal = if_else(year == 2021,
          true = 186,false = 155)) %>%
  mutate(prop2022 = totalrespondents / yeartotal) %>%
  filter(classyear != "Prefer not to state") %>%
  select(1, 5)
## create the table for proportion of each
## level of 2022 respondents for the class year question

survey_full %>% filter(year == 2021) %>% group_by(classyear, year) %>%
  summarise(totalrespondents = n()) %>%
  mutate(yeartotal = if_else(year == 2021,
                              true = 186,
                              false = 155)) %>%
  filter(classyear != "Prefer not to state") %>%
  mutate(prop2021 = totalrespondents/yeartotal) %>%
  select(1, 5) %>% ungroup() %>% mutate(prop2022 = classtable22$prop2022) %>%
  kable(caption = "Class Year Demographics", digits = 2)
## create the column for proportion of each
## level of 2021 respondents for the class year question
## and add that to the 2022 proportions
library(tidyverse)
```

```
survey_full <- survey_full %>% mutate(year = fct_relevel(year, c("2022", "2021"))) %>%
  filter(filling_frequency != "",
         filling_frequency != "Prefer not to state")

## relevel the year variable, and filter out blank responses or where
## they preferred not to state their response

## creating a plot colored by year, that shows the proportion of each factor level
## for daily water consumption
(ggplot(data = survey_full, aes(x = daily_water, y = after_stat(prop),
                                fill = year, group = year)) +
  geom_bar(position = "dodge") +
  coord_flip() +
  scale_fill_brewer(palette = "Set1", breaks = c("2021", "2022")) +
  theme_minimal(base_size = 14) +
  labs(title = "Daily Water Consumption",
       x = "Total Water Consumed",
       y = "Proportion of Respondents",
       colour = "Test") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12)))
library(tidyverse)
survey_full <- survey_full %>% mutate(year = fct_relevel(year, c("2022", "2021"))) %>%
  filter(filling_frequency != "",
         filling_frequency != "Prefer not to state")
## filter out "" values
## drop that level from the factor after filtering
## also get rid of Prefer not to state and then add in caption

ggplot(data = survey_full, aes(x = filling_frequency, y = after_stat(prop),
                               fill = year, group = year)) +
  geom_bar(position = "dodge") +
  coord_flip() +
  scale_fill_brewer(palette = "Set1", breaks = c("2021", "2022")) +
  theme_minimal(base_size = 14) +
  labs(title = "Frequency of Filling Station Usage",
       x = "Frequency",
       y = "Proportion of Respondents",
       colour = "Test") +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        legend.position = "bottom")
## creating a plot colored by year, that shows the proportion of each factor level
## for how often they used the filling stations
library(gridExtra)

survey_full <- survey_full %>%
  mutate(year = fct_relevel(year, c("2022", "2021"))) %>%
  filter(chemically_pure != "",
         safer_to_drink != "",
         clean != "")
## relevel the year variable and filter out blank responses
```

```
library(forcats)
survey_full$chemically_pure <- fct_collapse(survey_full$chemically_pure,
                                            Agree = c("Somewhat agree",
                                                      "Strongly agree"))
## group together somewhat agree and strongly agree into an agree level

survey_full$chemically_pure <- fct_collapse(survey_full$chemically_pure,
                                            Disagree = c("Somewhat disagree",
                                                         "Strongly disagree"))
## group together somewhat disagree and strongly disagree into a disagree level

chempureplot <- ggplot(data = survey_full, aes(x = chemically_pure, y = after_stat(prop),
                               fill = year, group = year)) +
  geom_bar(position = "dodge") +
  coord_flip() +
  scale_fill_brewer(palette = "Set1", breaks = c("2021", "2022")) +
  labs(title = "Water from the Filling Stations is more Chemically Pure",
       x = "Level of Agreement",
       y = "Proportion of Respondents") +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.y = element_text(size = 12))
## create a plot colored by year that shows the proportion of each factor level
## for the extent to which they agreed that the water from the filling
## stations is more chemically pure than tap water

survey_full$safer_to_drink <- fct_collapse(survey_full$safer_to_drink,
                                           Agree = c("Somewhat agree",
                                                     "Strongly agree"))
## group together somewhat agree and strongly agree into an agree level

survey_full$safer_to_drink <- fct_collapse(survey_full$safer_to_drink,
                                           Disagree = c("Somewhat disagree",
                                                        "Strongly disagree"))
## group together somewhat disagree and strongly disagree into a disagree level

safertodrinkplot <- ggplot(data = survey_full,
                           aes(x = safer_to_drink, y = after_stat(prop),
                               fill = year, group = year)) +
  geom_bar(position = "dodge") +
  coord_flip() +
  scale_fill_brewer(palette = "Set1", breaks = c("2021", "2022")) +
  labs(title = "Water from the Filling Stations is Safer to Drink",
       x = "Level of Agreement",
       y = "Proportion of Respondents") +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.y = element_text(size = 12))
## create a plot colored by year that shows the proportion of each factor level
## for the extent to which they agreed that the water from the filling
## stations is safer to drink than tap water
```

```
survey_full$clean <- fct_collapse(survey_full$clean,
                                  Agree = c("Somewhat agree",
                                            "Strongly agree"))
## group together somewhat agree and strongly agree into an agree level

survey_full$clean <- fct_collapse(survey_full$clean,
                                  Disagree = c("Somewhat disagree",
                                               "Strongly disagree"))
## group together somewhat disagree and strongly disagree into a disagree level

cleanplot <- ggplot(data = survey_full, aes(x = clean, y = after_stat(prop),
                               fill = year, group = year)) +
  geom_bar(position = "dodge") +
  coord_flip() +
  scale_fill_brewer(palette = "Set1", breaks = c("2021", "2022")) +
  labs(title = "The Filling Stations are Clean",
       x = "Level of Agreement",
       y = "Proportion of Respondents") +
  theme_minimal(base_size = 12) +
  theme(axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12),
        axis.title.y = element_text(size = 12))
## create a plot colored by year that shows the proportion of each factor level
## for the extent to which they agreed that the filling stations are clean

agreementplots <- grid.arrange(chempureplot, safertodrinkplot, cleanplot)
## arrange all three plots together
load(file = here("data/location_df21.rda"))
load(file = here("data/location_df22.rda"))
## load in the location dataframes

location_df21 <- location_df21 %>%
  mutate(
    location = fct_recode(
      location,
      `Student Center - 1st Floor` = "sc1",
      `ODY - Main Floor` = "odymain",
      Newell = "newell",
      `Fox Hall Lobby` = "madilllobby",
      Sykes = "sykes"
    ))
## recode some of the locations from 2021

location_df22 <-
  location_df22 %>% mutate(
    location = fct_recode(
      location,
      `Student Center - 1st Floor` = "sc1",
      `ODY - Main Floor` = "odymain",
      Newell = "newell",
      `Student Center - 3rd Floor` = "sc3",
      Whitman = "whitman"
    ))
```

```r
## recode some of the locations from 2022

location_df22_top <- location_df22 %>%
  arrange(desc(totalrespondents)) %>%
  ungroup %>% slice(1:5)
## create a dataframe with the top five most frequently used locations in 2022

## creating a plot to see which locations are used the most often in '22
locationplot22 <- ggplot(data = location_df22_top, aes(x = fct_reorder(location, totalrespondents), y =
  geom_point(size = 5) +
  geom_segment(aes(x=location, xend=location, y=0, yend=prop), size = 1.5) +
  coord_flip() +
  labs(y = "Proportion of Respondents", x = "Location") +
  theme_minimal(base_size = 13) +
  ylim(0, 0.9)

location_df21_top <- location_df21 %>%
  arrange(desc(totalrespondents)) %>%
  ungroup() %>% slice(1:5)
## create a dataframe with the top five most frequently used locations in 2021

## creating a plot to see which locations are used the most often in '21
locationplot21 <- ggplot(data = location_df21_top, aes(x = fct_reorder(location, totalrespondents), y =
  geom_point(size = 5) +
  geom_segment(aes(x=location, xend=location, y=0, yend=prop), size = 1.5) +
  coord_flip() +
  labs(y = "Proportion of Respondents", x = "Location") +
  theme_minimal(base_size = 13) +
  ylim(0, 0.9)

locationplotfull <- grid.arrange(locationplot21, locationplot22)
library(tidyverse)
library(here)
load(here("data/wc_long_iter_total.rda"))
wc_long_iter_total <- wc_long_iter_total %>% arrange(desc(proptimes))
## arrange to see which teams made it out of the group stages
## the highest proportion of times

actually_went_through_18 <- c("France", "Argentina", "Uruguay", "Portugal",
                              "Brazil", "Mexico", "Belgium", "Japan",
                              "Spain", "Russia", "Croatia", "Denmark",
                              "Sweden", "Switzerland", "Colombia", "England")
## create a list of teams that actually made it through in 2018

wc_long_18_through <- wc_long_iter_total %>% filter(team %in% actually_went_through_18)
## create a data frame to filter out the teams that actually made it through

(ggplot(data = wc_long_iter_total, aes(x = proptimes, y = fct_reorder(team, proptimes))) +
  geom_point() +
  geom_segment(data = wc_long_iter_total, aes(x = 0, xend = proptimes, y = team, yend = team)) +
  geom_segment(data = wc_long_18_through, aes(x = 0, xend = proptimes, y = team, yend = team), colour =
  geom_point(data = wc_long_18_through, aes(x = proptimes, y = team), colour = "red") +
  facet_wrap(~group, scales = "free_y") +
  theme_minimal() +
  labs(x = "Proportion of Times a Team Advanced out of the Group Stages",
       y = "Team",
      title = "2018 World Cup Group Stage Predictions"))
## create a plot with the proportion of times a team advanced out of the group stages
## with the teams in red being the teams who actually made it out of the group stages
library(tidyverse)
library(here)
load(here("data/wc_long_iter_total_22.rda"))
actually_went_through <- c("Netherlands", "USA", "Argentina", "Australia",
                           "Japan", "Croatia", "Brazil", "South Korea",
                           "England", "Senegal", "France", "Poland",
                           "Morocco", "Spain", "Portugal", "Switzerland")
## create a list of teams that actually made it through in 2022

wc_long_22_through <- wc_long_iter_total_22 %>% filter(team %in% actually_went_through)
## create a data frame to filter out the teams that actually made it through

wc_long_iter_total_22 <- wc_long_iter_total_22 %>% arrange(desc(proptimes))
## arrange to see which teams made it out of the group stages
## the highest proportion of times

(ggplot(data = wc_long_iter_total_22, aes(x = proptimes, y = fct_reorder(team, proptimes))) +
  geom_point() +
  geom_segment(data = wc_long_iter_total_22, aes(x = 0, xend = proptimes, y = team, yend = team)) +
  geom_segment(data = wc_long_22_through, aes(x = 0, xend = proptimes, y = team, yend = team), colour =
  geom_point(data = wc_long_22_through, aes(x = proptimes, y = team), colour = "red") +
  facet_wrap(~group, scales = "free_y") +
  theme_minimal() +
  labs(x = "Proportion of Times a Team Advanced out of the Group Stages",
       y = "Team",
       title = "2022 World Cup Group Stage Predictions"))
## create a plot with the proportion of times a team advanced out of the group stages
## with the teams in red being the teams who actually made it out of the group stages
```