Kylie Hoar, Joanna Hu, Mirabelle El Chalfoun, and Josiah Aranovitch
12/11/2023

DS 2010 Final Project Report

A Water Potability Analysis of Virginia Beach, Virginia

Data Set Background

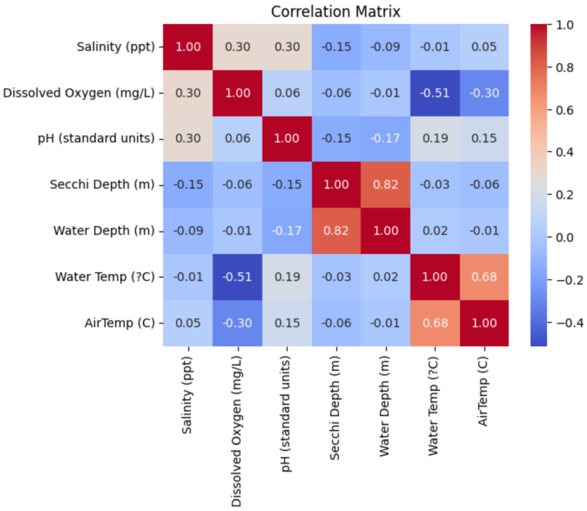
Our data set analyzes the water quality data collected from various water bodies in Virginia Beach, such as the Bay, A-pool, B-pool, C-pool, D-pool. The variables we looked at are salinity, dissolved oxygen, pH, secchi depth, water depth, water temperature, time, date, and air temperature. The data set website states that high salinity, low dissolved oxygen, high or low pH, and high secchi depth are important to conserve high water quality standards so we focused on those variables. Maintaining good water quality is important as it keeps aquatic vegetation alive and the food chain in a healthy state.

Motivation

We decided to pursue this project due to our shared passion for environmental concerns and wildlife. We thought it would be interesting to look at how water potability changes over time, as that can give us answers to questions about the health of the bay and can even tell us what the ecosystem is like in Virginia Beach. By looking at how certain variables fluctuate over time, such as water temperature, we can get an idea of how climate change has affected the Back Bay and Virginia Beach overall. As advocates for environmental conservation, we hope that our project contributes valuable data that not only promotes sustainable practices but also serves as a resource to educate the community about the importance of climate change and taking care of your local beaches. Through our efforts, we hope to have inspired a sense of responsibility for the well-being of Back Bay and other beaches.

Kylie Hoar, Joanna Hu, Mirabelle El Chalfoun, and Josiah Aranovitch
12/11/2023

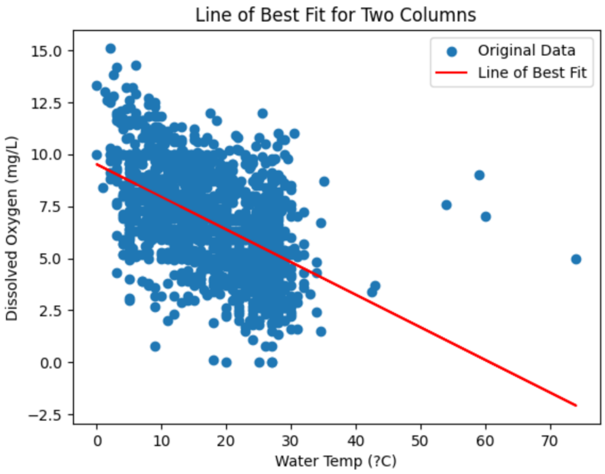
Conjecture 1: Air Temperature vs Water Temperature vs Dissolved Oxygen

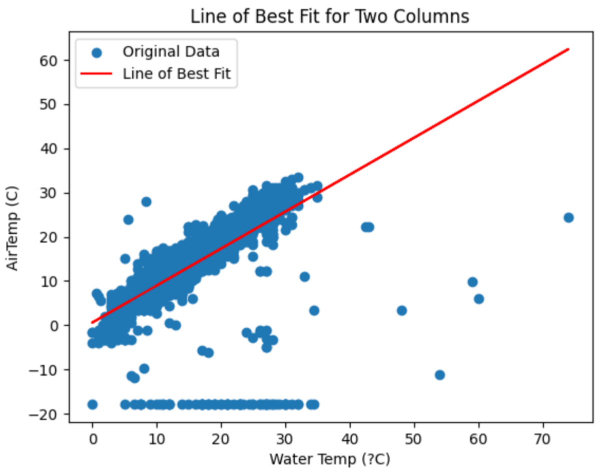


	Salinity (ppt)	Dissolved Oxygen (mg/L)	pH (standard units)	Secchi Depth (m)	Water Depth (m)	Water Temp (?C)	AirTemp (C)
Salinity (ppt)	1.00	0.30	0.30	-0.15	-0.09	-0.01	0.05
Dissolved Oxygen (mg/L)	0.30	1.00	0.06	-0.06	-0.01	-0.51	-0.30
pH (standard units)	0.30	0.06	1.00	-0.15	-0.17	0.19	0.15
Secchi Depth (m)	-0.15	-0.06	-0.15	1.00	0.82	-0.03	-0.06
Water Depth (m)	-0.09	-0.01	-0.17	0.82	1.00	0.02	-0.01
Water Temp (?C)	-0.01	-0.51	0.19	-0.03	0.02	1.00	0.68
AirTemp (C)	0.05	-0.30	0.15	-0.06	-0.01	0.68	1.00

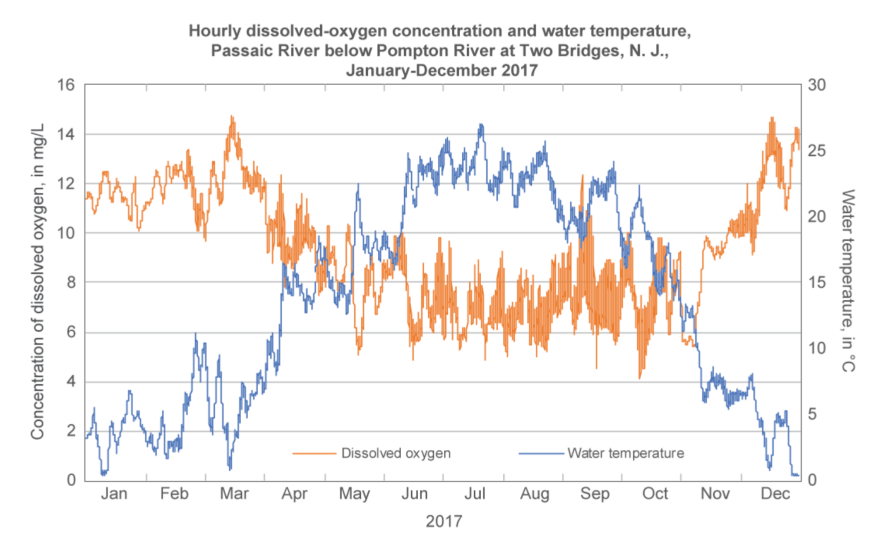
Based on the correlation matrix, there is a -0.51 correlation between dissolved oxygen levels (mg/L) and water temperature (C). This means that when the water temperature rises, the oxygen levels decrease. Also, water temperature (C) and air temperature (C) have a 0.68 correlation, which means that as air temperature increases, so does water temperature. It is important to know that correlation does not necessarily imply causation.

In the visualizations below, the relationship between each variable is represented with a line of best fit based on the dataset. Again, they both show that water temperature is inversely related to dissolved oxygen levels, and water temperature is directly related to air temperature. Therefore, based on this dataset, it seems that as temperatures increase, the dissolved oxygen levels decrease in water, which results in lower water potability.





According to the United States Geological Survey (2018), cold water is able to hold onto oxygen better than warm water, so the oxygen levels in water are higher in winter and early spring when it is colder and lower in summer and fall when the temperatures are higher. As a result, this conjecture is found to be true.



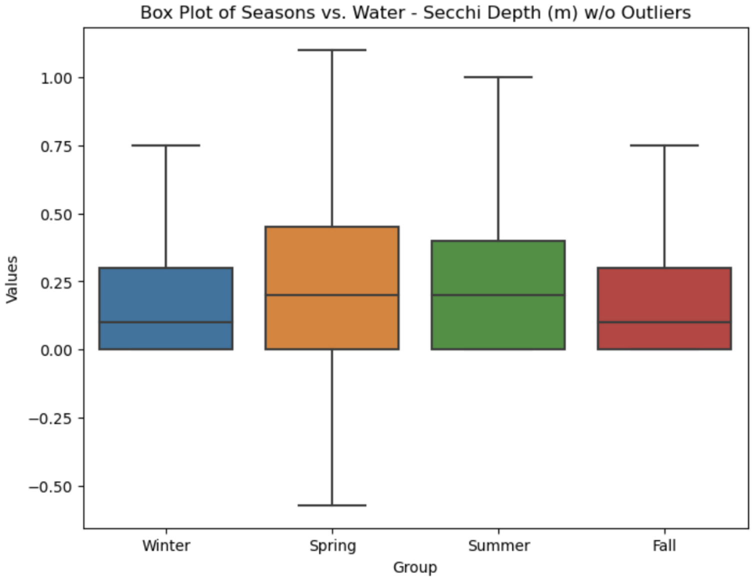
Conjecture 2: Spring and Summer are the Seasons With the Worst Water Potability

Our team determined that Spring and Summer would be the seasons with the worst water potability because of an article published by Tom Glanville at Iowa State University. The author concluded that the best seasons to test well water would be Spring or Summer due to buildup of bacteria in water and heavy rainfall. With all of these potential pollutants in the water, we determined that Spring and Summer would be the seasons with the lowest water potability (Glanville).

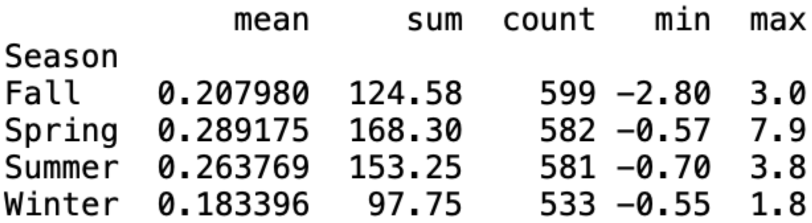
To measure water potability across different seasons, we decided to measure the difference between Water Depth and Secchi Depth. Secchi Depth is the measure of water transparency. To measure Secchi Depth, a disk is inserted into the water and once it is no longer visible a measure of the depth of the disk is taken. Secchi Depth directly relates to water turbidity, a defining factor of water potability, or safety. High turbidity or a low Secchi Depth correlates to low water potability and, conversely, a low turbidity or a high Secchi Depth correlates to high water potability. The reason why we subtracted Secchi Depth from Water Depth in our analysis of seasonal changes in water potability is to ensure that our comparisons were not affected by the depth of the samples taken. Additionally, we wanted to focus on the differences in Secchi Depth and Water Depth to highlight how much turbidity is in the water. In this situation, the greater the difference between Water Depth and Secchi Depth, the more turbid and less safe the water will be. Conversely, the lesser the difference between Water Depth and Secchi Depth, the less turbid and more safe the water will be.

For some context on our distribution of seasons, Winter is defined by the months of December, January, and February; Spring is defined by the months of March, April, and May; Summer is defined by the months of June, July, and August; and Fall is defined by the months of September, October, and November.

In our analysis of the Virginia Beach water potability dataset, the following boxplot was created to compare how seasons would affect the measure of the difference between Water Depth and Secchi Depth in meters.



In the visualization of these boxplots, it is apparent that Winter has the smallest difference between Water and Secchi Depth on average, and is therefore the time of year with the highest water potability. In contrast, it seems like Spring has the greatest difference between Water and Secchi Depth on average, and is therefore the time of year with the lowest water potability. In addition to the visualization, a statistical summary of these plots is shown below and proves that Winter is in fact the worst time of year in terms of water potability and Spring is the best.

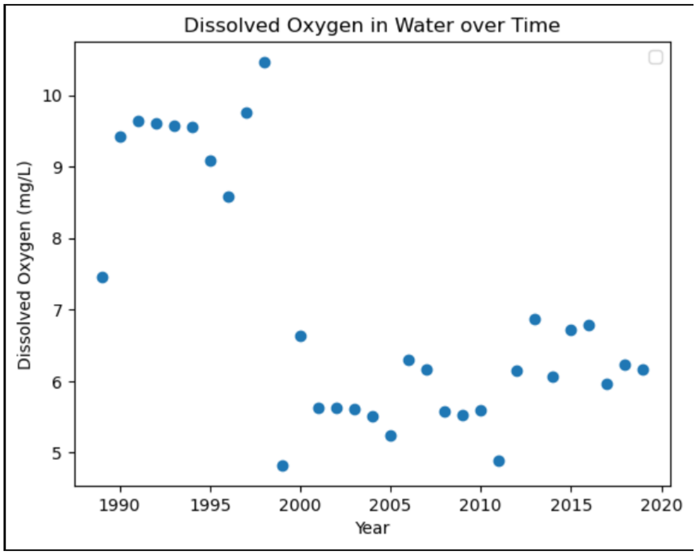


```
            mean     sum  count   min  max
Season
Fall    0.207980  124.58    599 -2.80  3.0
Spring  0.289175  168.30    582 -0.57  7.9
Summer  0.263769  153.25    581 -0.70  3.8
Winter  0.183396   97.75    533 -0.55  1.8
```

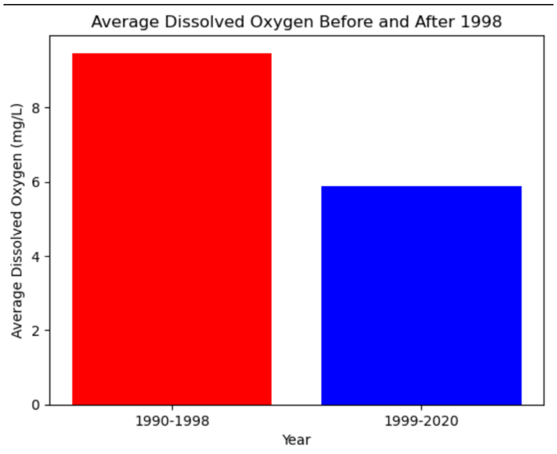
Overall, our conjecture that Spring is the season with the lowest water potability is true. Spring has the largest difference between Water and Secchi Depth on average.

Conjecture 3: Average Dissolved Oxygen Levels in Water Decreases over Time

As observed in the data taken from Virginia Beach, oxygen levels in the 1990s consistently hovered around 9 mg/L. Then, at the end of the decade, a sudden drop to about 6 mg/L ensued. Since then, the oxygen level in the waters have never rebounded, prompting an investigation. Marine environmentalists at Yale University have noted that "Oxygen levels in the world's oceans have already dropped more than 2 percent between 1960 and 2010, and they are expected to decline up to 7 percent below the 1960 level over the next century" (Jones, 2023). The cause of this is said to be global warming. It is therefore not surprising that water oxygen levels have decreased from 1990 to 2020. However, no information was found to suggest why there was such a drastic drop in oxygen levels between 1998 and 1999.



A drop in the oxygen level by a couple percentage points would suggest that global warming is the driving factor behind the water oxygen level in Virginia Beach. But the average oxygen levels from 1999-2020 represents a 37.9% drop from 1990-1998 averages, as seen from the figure to the right. As there was no information found online to prove this large drop in oxygen happened, it is possible that the data may be not fully correct. It could also mean an event that impacted oxygen levels did happen but is not well-documented. Our conjecture is true in that oxygen level decreases over time, with the caveat that



```
The average value between 1990 and 1998 is: 9.46841121495327
The average value between 1999 and 2020 is: 5.8765261382799325
```

there is only a noteworthy change in oxygen level over one year rather than over 30 years.

Classification Model

Our team decided to test three classification models on our water potability dataset. More specifically, we decided to focus on Conjecture #2 to see if we would be able to predict the season that a sample was collected in based on the Water - Secchi Depth (m) calculation. We conducted a Logistic Regression, KNN, and Random Forest Model. We obtained accuracy scores of 0.2614, 0.2135, and 0.2222 respectively. From all three of these values, we are able to conclude that the Water-Secchi Depth (m) calculation is a poor value to use to determine the season that a sample was collected in. Additionally, Logistic Regression, KNN, and Random Forest are poor classification models for this relationship between Water - Secchi Depth (m) and Season. A visualization of the accuracy scores from our model can be seen below.

Challenges Faced

1) One of the challenges we faced was finding a conjecture that had good accuracies when using a classification/regression model. We attempted to solve this problem by creating classification/regression models for each conjecture, but our best one only ended up having accuracies of around 0.24, which we ended up using.
2) Some of our members did not have much experience with coding yet, so it was more difficult for them when it came to the coding part. Luckily, some of our teammates had more experience with coding and were able to help those who needed it.

Contributions:

Mirabelle: Data set background (report and slides), Motivation (report and slides)
I worked on writing the data set background and motivation sections of the report and the slide. I also worked on the conclusion/project summary slide. I helped with the coding sections and gave my input whenever I could and created a correlation matrix to analyze the data. I also looked for correlations between other variables that were not included in our conjectures, but found no correlations so we did not end up putting them anywhere in our report or slides.

Joanna: Conjecture 1 (report and slides), Creating visualizations using code, compiling final Jupyter notebook
I worked on analyzing and creating visualizations for Conjecture 1 (Air Temp vs Water Temp vs Dissolved Oxygen Levels). I also helped in creating a heatmap for the correlation matrix so that it would be easier to look at and understand. I created a line of best fit visualizations comparing air temperature, water temperature, and dissolved oxygen levels to use in my sections of the slides. I also wrote the Conjecture 1 section in the report and contributed to any other necessary sections such as Challenges Faced and Contributions. I helped compile everyone's code into one Jupyter notebook. Finally, I also tried my best to help my teammates whenever I could.

Kylie: Conjecture 2 (report and slides), Classification Model (report and slides), Code Visualization
I worked on analyzing and creating visualizations for Conjecture 2 (Spring is the Season With the Worst Water Potability). I created statistical analyses of the boxplot created in Conjecture 2 to help for summarization and clarification. I wrote the Conjecture 2 part of the report, and designed the slides for the presentation. In addition to Conjecture 2, I created the Classification Model to analyze Conjecture 2. I wrote the report for this section and designed the slides for it as well. I also helped in creating a correlation matrix for the dataset to help understand the dataset better. Finally, I created our team group chat, sent when2meets to coordinate team meeting times, and assisted teammates whenever I could.

Josiah: Conjecture 3 (report and slides), Creating visualizations using code
I worked on visualization and analysis for Conjecture 3 (Average Dissolved Oxygen Levels in Water Decreases over Time). I made a scatter plot to show how oxygen

levels in the water change over time. I also produced a simple bar chart that compared the average yearly oxygen levels for pre-1998 and post-1998 against each other to visualize a large change in oxygen that happened that year. Both of these visualizations can be found in the Conjecture 3 section that I wrote. Finally, I assisted with data cleaning to make our dataset less complicated.

References

Dissolved oxygen and water completed. Dissolved Oxygen and Water | U.S. Geological Survey. (2018). https://www.usgs.gov/special-topics/water-science-school/science/dissolved-oxygen-and-water.

Glanville, T. (n.d.). *Spring is a good time to test well water*. News. https://www.extension.iastate.edu/news/2009/may/110401.htm.

Publisher U.S. Fish and Wildlife Service. (2023, October 29). *Water Quality Data.* Catalog. https://catalog.data.gov/dataset/water-quality-data-41c5e

Jones, N. (2023, May 11). As *ocean oxygen levels dip, fish face an uncertain future*. Yale E360. https://e360.yale.edu/features/as-ocean-oxygen-levels-dip-fish-face-an-uncertain-future#:~:text=Oxygen%20levels%20in%20the%20world's,15%20percent%20of%20its%20oxygen.

Becker, K. (2016, January 29). *Understanding Dissolved Oxygen*. GrowerTalks. https://www.growertalks.com/Article/?articleid=22058