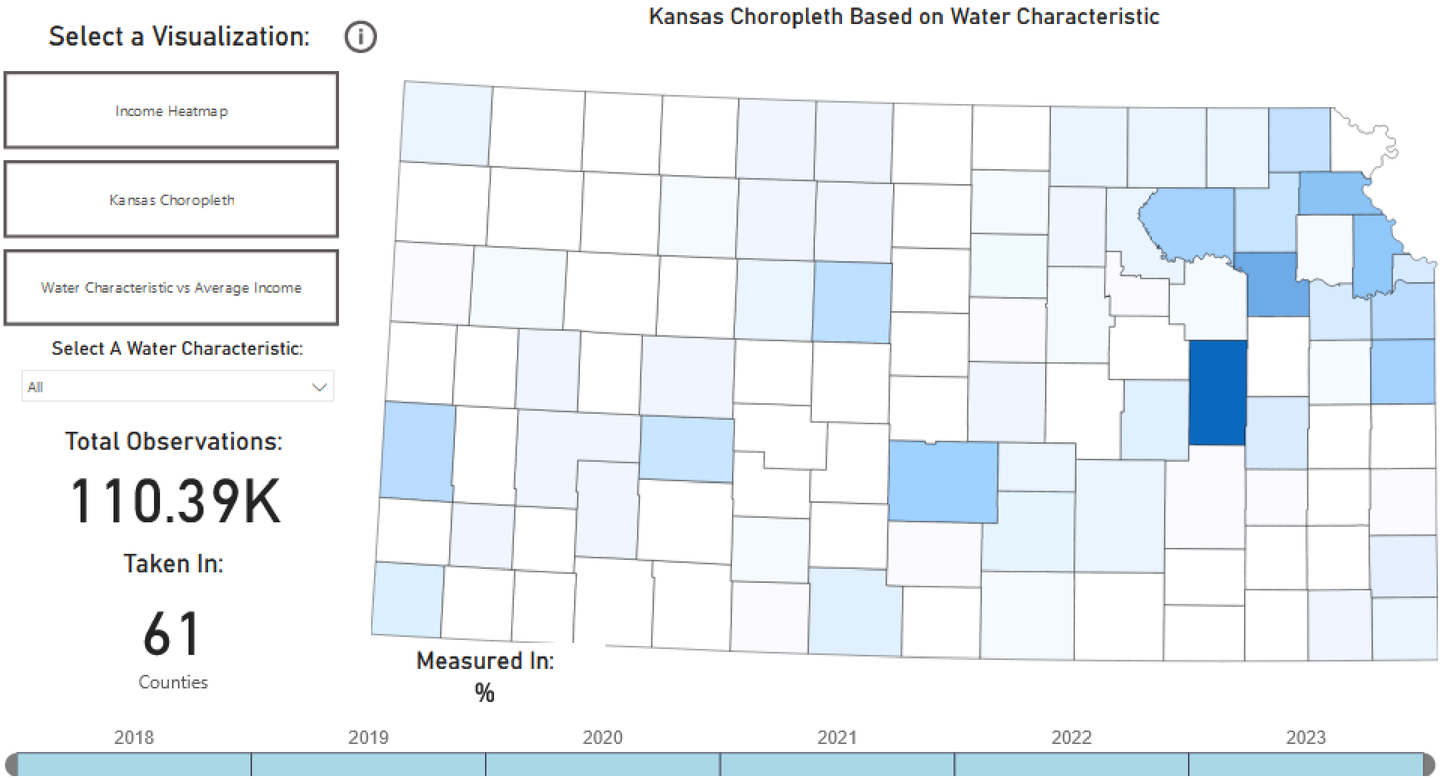
Water Quality Portal Dashboard

Kansas Water Quality Insights: A Socio-Economic Perspective

Caro Champion, Gregory Shoda, Erik Kreider, Muhammad Siddiq Khan, Melanie Erkman


Select a Visualization:
Income Heatmap
Kansas Choropleth
Water Characteristic vs Average Income
Kansas Choropleth Based on Water Characteristic
2018
2019
2020
2021
2022
2023

Select A Water Characteristic: All

Total Observations: 110.39K

Taken In: 61 Counties

Measured In: %

Fig. 1. Kansas Choropleth Visualization from our Water Quality Dashboard [1]

Introduction

The Water Quality Portal (WQP) is a collaborative effort between the United States Geological Survey (USGS) and the Environmental Protection Agency (EPA). It provides access to water quality data collected by government agencies across the United States. The portal allows users to filter and then export datasets which can be used for in-depth analysis.

Understanding water quality in local areas can be useful for academia, local agencies, commercial developers, and the general public. Academics could identify the causes behind trends in water quality in order to either correct or replicate said trends. Local agencies must monitor water quality to avoid health risks to the local populace and wildlife. Commercial developers would want to understand water quality conditions when looking to build living accommodations, as it could negatively affect property value and earning potential. Finally, the general public might consider water quality and its impact on health before moving to a new location.

Project Objectives

We were tasked with creating a dashboard that allows users to visualize and analyze Kansas water quality data. Our goal was to create an interactive, easy-to-use tool that informs users about different metrics of Kansas' water quality and relate those qualities to meaningful socio-economic measures. WQP data, in its raw form, provides site-specific location data, so we wanted our dashboard to use aggregated, county-level data to reduce visual clutter and provide practical insights to users.

Related Work

Published research related to the Water Quality Portal (WQP) introduces the databases' scope, purpose, and possible uses (Read et. al. 2017) [2]. This report highlights the goal of the WQP, which is to serve as "a single point of access for national-scale water quality data to facilitate analysis and decision making at the local, regional, and national scales." The published report shared several visualizations, including a geospatial choropleth map, which details the density of water sites in the US. It also provided a stacked bar chart to summarize the types of water sites available and a line chart to illustrate seasonal changes in water depth throughout the US over time.

Additionally, a Water Quality Portal Dashboard was published by the Coastal & Heartland National Estuary Partnership in collaboration with the University of South

Florida Water Institute (2018) [3]. This dashboard was made using ArcGIS and focuses on 5 key metrics recorded by researchers at water sites: chlorophyll a, phosphorus, nitrogen, Secchi depth, and bacteria (*E. coli./Enterococcus.*) The dashboard's goal is to display water conditions in the Tampa Bay area based on the most recently recorded water quality sample per site location. A map of the Tampa Bay area is the primary visualization and water quality is indicated by color in three tiers: good, fair, and poor. Water quality thresholds are dictated by the Florida Administrative Code and implemented directly into the supporting dataset.

Data

For our project, we used a data set created by combining data from the Water Quality Portal, data from the IRS Tax Statistics, and basic geographic data (2023; 2021) [4][5]. The Water Quality Portal allows CSV exports of water quality sampling across test sites in various States and multiple decades. Due to the scope of the project, we filtered down collection sites to those located in Kansas, and due to dataset size limitations, we also filtered down the data to entries collected from 2018 to 2023. We used Microsoft Excel to combine our water quality data exports with the IRS income data based off of county name.We pared down this filtered data set even further by removing approximately 100 fields. This left us with 33 columns and 173,140 rows. Each row represents a recorded water quality measurement, and each column represents attributes relating to that data record.

Some of the most important attributes describe the *who, when,* and *where* of water quality collection: the organization identifier, the type of media used for collection (surface or groundwater), the collection start date, longitude, latitude, city, county, state, and ZIP code. Other major attributes describe the type of characteristic being measured, its measured value, and the relevant units of measurement. Some examples of characteristics include pH level, temperature, phosphorus levels, and stream flow. We also chose average gross income by county as the socioeconomic attribute to focus on.

According to some simple statistics generated from the dataset, the earliest recorded measurement was taken on January 8th, 2018, and the most recent measurement was taken on December 21st, 2023. Out of the 173,140 data records, 74,214 were missing measured values. Due to the size of the dataset we decided to simply remove null values. Substituting in average or median values in this dataset could potentially skew the data and paint a picture that the measurements don't truly represent. In terms of average gross income by county, the minimum was $950 from Woodson County, the median was $35,208 from Geary County, and the maximum was $218,289 from Johnson County.

Visualizations

We created our dashboard in PowerBI. Our dashboard provides our users with three key visualizations: 1) a choropleth map that shows average income by county in 2021; 2) another choropleth map that shows county-wide averages of a selected water characteristic; and 3) a scatterplot that visualizes the correlation between average income and average water characteristic measure by county [6][7]. Our dashboard is equipped with a zoom functionality so users can magnify any visualization to better suit their viewing needs. Additionally, the user can click the "Information" button (depicted as a circle "i") at any time to be directed to this report.

The income choropleth shows the average income by county using a red, monochromatic gradient. Deep red indicates a high average income, and conversely, white indicates a low average income or a lack of data. Users can hover over the county of their choice to get its name and its exact average income. Additionally, users can click the "Focus Mode" button at the bottom left-hand corner of the map to isolate the map itself and remove the other visualization options.

The water characteristic choropleth shows the average measurement of a selected water quality characteristic by county using a blue, monochromatic gradient. Deep blue indicates a high average measure, while white indicates a low average measure or a lack of data. There is a slicer on the left that allows users to select a water quality characteristic for visualization and a timeline-based selection tool on the bottom that allows users to select the range date of their data. When users make their selections, the data summary card on the left will update to show the total number of observations for a given water characteristic in a given time period, as well as the number of counties the characteristic was measured in. Like with the income choropleth, users can hover over a county for more details or use the "Focus mode" to isolate the map itself.

Our last visualization plots county-level average incomes against county-level water characteristic averages. The scatter plot comes with a linear regression line that depicts the average relationship between the two quantities. Additionally, the correlation coefficient is provided in the top right corner of the visualization. As with our previous visualization, we can select a water quality characteristic and a time period of interest. These selections will be reflected in both the summary card to the left and the scatter plot in the center.

Key Insights

The first visualization in the water quality dashboard shows how average income is distributed across the state of Kansas. We can see that in general, the highest income counties are located in the upper-right corner of the state, and the average county-level income drops off as we get further away. We are missing some income information for certain counties, so other "pockets" of high-income counties may exist elsewhere.

There are 50 water quality characteristics that can be visualized in either the choropleth or the scatter plot. This means that there are a lot of characteristic-specific insights that can be gained depending on what the user is most interested in. The availability and distribution of data depends heavily on the chosen characteristic. Additionally, with how the water quality data was collected, there are often counties with no measurements for specific characteristics.

Discussion

When we started our report, we had not finalized our dataset. Unsurprisingly, the feedback we received afterwards asked for specific details about our dataset that we didn't have at the time. During our redesign, we went back and introduced some specifics of our dataset in the beginning of the report. That way, we could open the paper by linking the purpose of our project directly to the capabilities of our dataset. With our redesign, we were able to also produce a dashboard in PowerBI that better matches up with the purpose of our water quality project. Unlike the initial mock-up and visualizations we provided, our end product is interactive and fulfills our project objectives.

To validate our report, we gathered additional feedback from people outside of the course. Their feedback was helpful in making sure that our report made sense to those who were unfamiliar with our project and/or had less of a technical background. To validate our dashboard, we did plenty of user testing within our group to make sure that it was functional and provided useful insights into water quality and income across Kansas' counties.

Challenges

Throughout the process of compiling, preprocessing and analyzing the Water Quality dataset, we were able to identify some key challenges and opportunities. Initially, we noticed an issue with the size of the dataset. Our original export of Kansas Water Quality data for "all-time" data immediately proved too large of a dataset for Excel to handle. This means we were left with 2 options: filter down the dataset to more recent years, or connect to the dataset via an API tool and store the data in a database. Due to the limited time available for this project, we chose the option to filter down the dataset, which ultimately limits the ability to compare water quality of sampled sites throughout the decades. A future opportunity related to this challenge would be to connect to the dataset via an API and store the entire dataset in a relational database. This would allow for easy analysis via Python, our programming language of choice.

Another issue we identified was around limited domain knowledge of the underlying dataset. We found it difficult to determine whether a characteristic was important to overall water quality, let alone if water quality was considered "contaminated" at the time of a sampling. With our dashboard, we opted to focus on the 50 characteristics with the most samples because it would give us the most data to work with. A future improvement would be to involve a domain expert. An expert could help guide the team in understanding key characteristics, critical levels of these characteristics, and the most impactful way to visualize these metrics for a key stakeholder.

Additionally, a key stakeholder would be able to identify potential socioeconomic variables that relate to our selected characteristics. We chose to focus on mean gross income and assumed that areas of higher wealth might invest more in water quality health. However, there are many more socioeconomic metrics which might be worth investigating, such as metro vs. urban, recent construction areas, zoning (e.g. commercial vs. residential), etc.

Acknowledgements

We would like to thank Dr. Andreas Bueckle, Michael Ginda, and our project sponsor Dr. William Duncan for giving us the opportunity to work on this project. Additionally, we would like to thank our peers and friends who gave us feedback on our report.

REFERENCES

1. Shoda, G. (2024, April 11). *Water Quality Portal Dashboard*. Power BI. https://app.powerbi.com/view?r=eyJrIjoiYzk5ZGFlMDItYzliMS00MWU4LWE3MmUtYjgyYzU1OTY0NzM0IiwidCI6IjExMTNiZTM0LWFlZDEtNGQwMC1hYjRiLWNkZDAyNTEwYmU5MSIsImMiOjN9
2. Read, E. K., Carr, L., Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., Kreft, J., Read, J. S., & Winslow, L. A. (2017, January 24). *Water quality data for national‐Scale Aquatic Research: The Water Quality Portal.* Water quality data for national-scale aquatic research: The Water Quality Portal. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1002/2016WR019993
3. USF Water Institute. (2018, June 1). *CHNEP Water Atlas*. Water Quality Dashboard - CHNEP.WaterAtlas.org. https://chnep.wateratlas.usf.edu/water-quality-dashboard
4. National Water Quality Monitoring Counsil. (n.d.-b). Water Quality Data Home. https://www.waterqualitydata.us/
5. Internal Revenue Service. (n.d.). *Soi tax stats - individual income tax statistics - 2021 ZIP code data (SOI)*. Individual Income Tax Statistics. https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2021-zip-code-data-soi-0
6. Upadhyay, A. (2023, September 29). -. Igismap. https://www.igismap.com/
7. Ben. (2021, October 29). *Correlation coefficient in power BI using Dax*. Ben's Blog. https://datakuity.com/2021/10/29/correlation-coefficient-in-power-bi-using-dax/