SMU | COX SCHOOL OF BUSINESS

MSBA Capstone Project Executive Summary

EXECUTIVE SUMMARY

The Capstone project for the City of Dallas focused on three key areas of development for City of Dallas Data Scientist Ash Nyangani to assist him in the analysis of water quality data across two Dallas watersheds. This included 1) Tableau visualizations of data from the Bachman Branch and Turtle Creek watersheds from 2009 through 2019 to help understand the water quality and pollutant levels of each watershed, 2) an R file with steps for efficient data cleaning and statistical analysis that can be used for data from all 30 remaining watersheds moving forward, and 3) a PowerPoint presentation summarizing the process of data cleaning and analysis as well as issues that arose along the way and logical next steps for the City to conduct water quality analysis moving forward.

PowerPoint Project:
The following questions should be considered when building the Project PowerPoint:

- How do we relay the process of cleaning the data and visualizing so that it is intuitive to the end user?
- What data gathering steps should be taken to make the cleaning process easier?
- What is the average level of water quality across watersheds?
- Are either one of the watersheds at risky pollutant levels?
- What steps should be taken next to replicate the project across the remaining 30 watersheds?

Visualizations in Tableau:
The following questions should be considered when building the visualizations in Tableau:

- What trends can be seen in each pollutant over the past decade?
- Which pollutants should be considered most important in assessing water quality?
- What is the average level for each of the elements?
- How do the pollutant levels differ across watersheds?
- How can the data be visualized geospatially across both watersheds?

R File:
The following questions should be considered when building the R code:

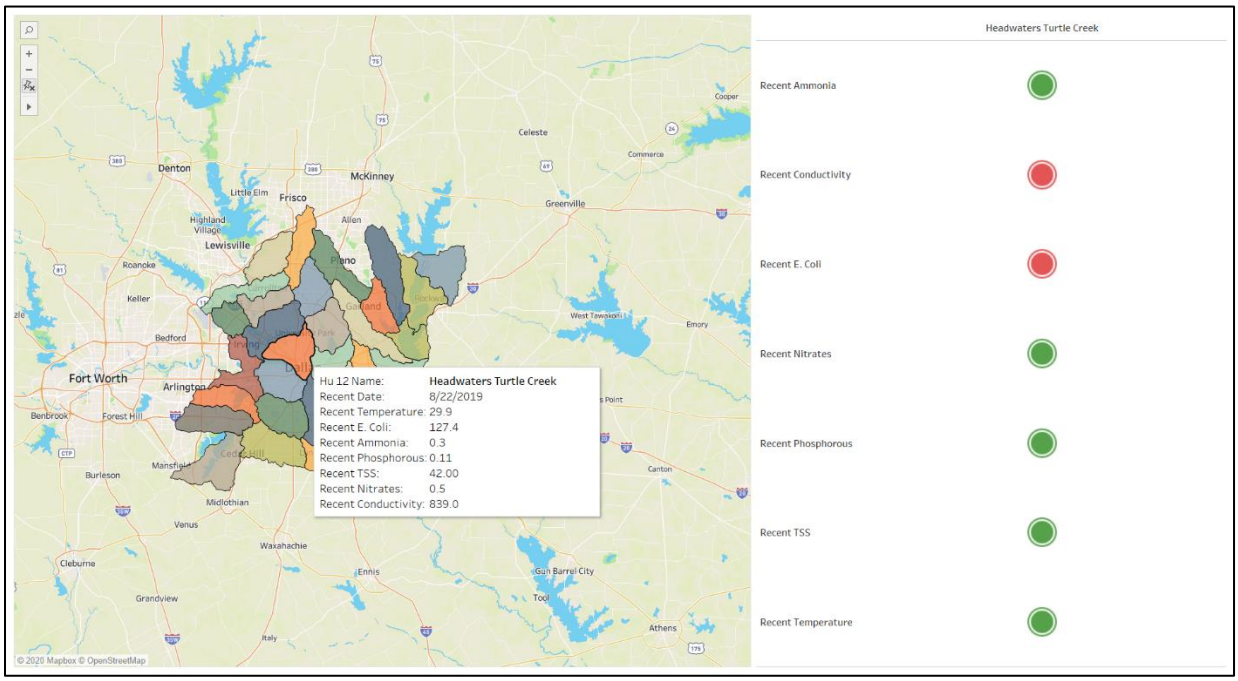
- How should strings be handled in the data?
- What should be done with “<” and “>” before numerical values in the data?
- How should merged cells be cleaned?
- Is the code easy to replicate for new watershed data?
- What statistics should be performed on the data to analyze water quality?

For the Tableau portion, five Dashboards were created to assist in answering these questions:

- Map with KPI Dashboard
- Watershed Drilldown Dashboard
- Watershed Comparison Dashboard
- Gauge Charts 1
- Gauge Charts 2

All acceptable parameters referred to in this project come from the Texas Surface Water Quality Standards and TPDES Benchmark values.

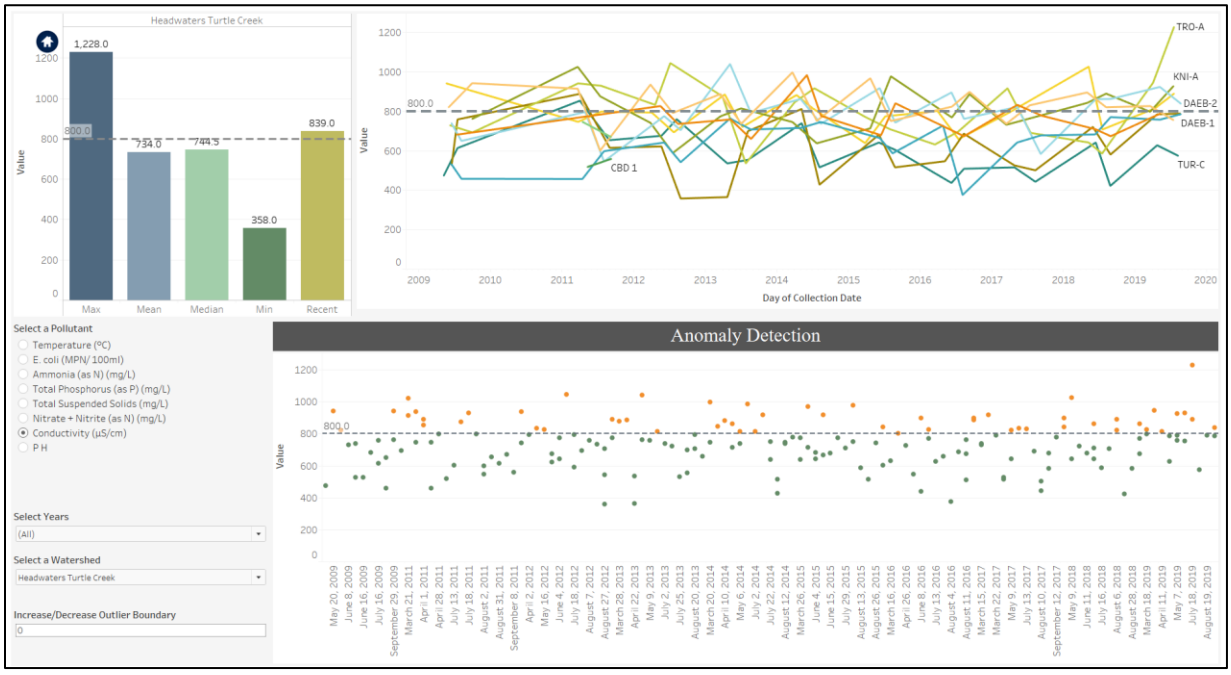
MAP WITH KPI



Business Insights:

- This dashboard allows the viewer to see the size and location of each watershed, as well as where watersheds are in comparison to each other. Once data is added for all watersheds, it would be possible to see if the quality of water in one location is affected by the watersheds that it neighbors. This would be helpful in deciding where and how to treat water in order to positively affect the quality of multiple watersheds.
- The right half of the dashboard shows a circle for each of the top eight pollutants, colored according to whether the most recent sampling data is inside its respective acceptable parameter. This element of the dashboard is intended to show the most current data points and how they are measuring up to the goals of the City.
    - In the most recent samples, both the Bachman Branch and Headwaters Turtle Creek branches are outside of the acceptable parameter for E. Coli. In addition, Headwaters Turtle Creek is outside of the Conductivity acceptable parameter.

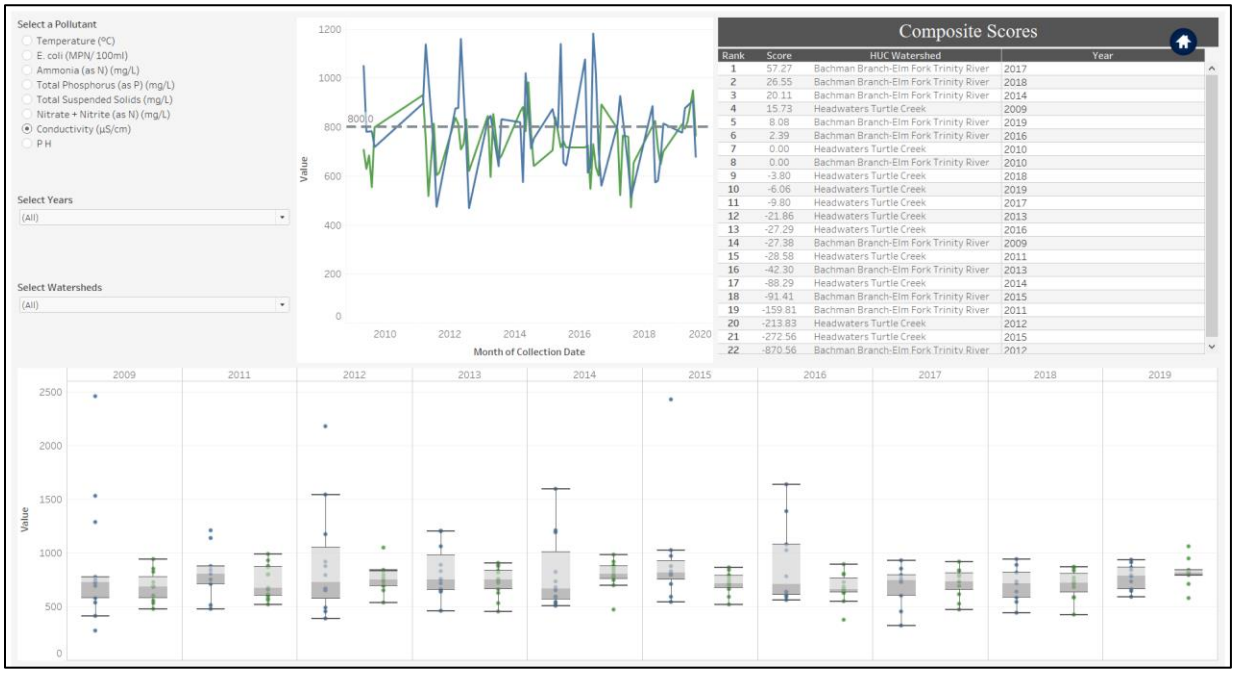
WATERSHED DRILLDOWN



Business Insights:

- All values of temperature, ammonia, phosphorus, and pH in the Headwaters Turtle Creek watershed have been inside their acceptable parameters for the past decade.
- The Bachman Branch has had values outside of the set parameters for all pollutants except Temperature.
- The pollutant with the most extreme outliers by far for both watersheds is E. Coli.
- The ELMT-1 and DAN-A sampling sites in the Bachman Branch watershed and the DAEB-1 and CBD-2 sampling sites in the Headwaters Turtle Creek watershed seem to have the most outliers.
- Nitrates, Conductivity, and E. Coli have the most outliers in Headwaters Turtle Creek, and E. Coli and Conductivity have the most outliers for the Bachman Branch.

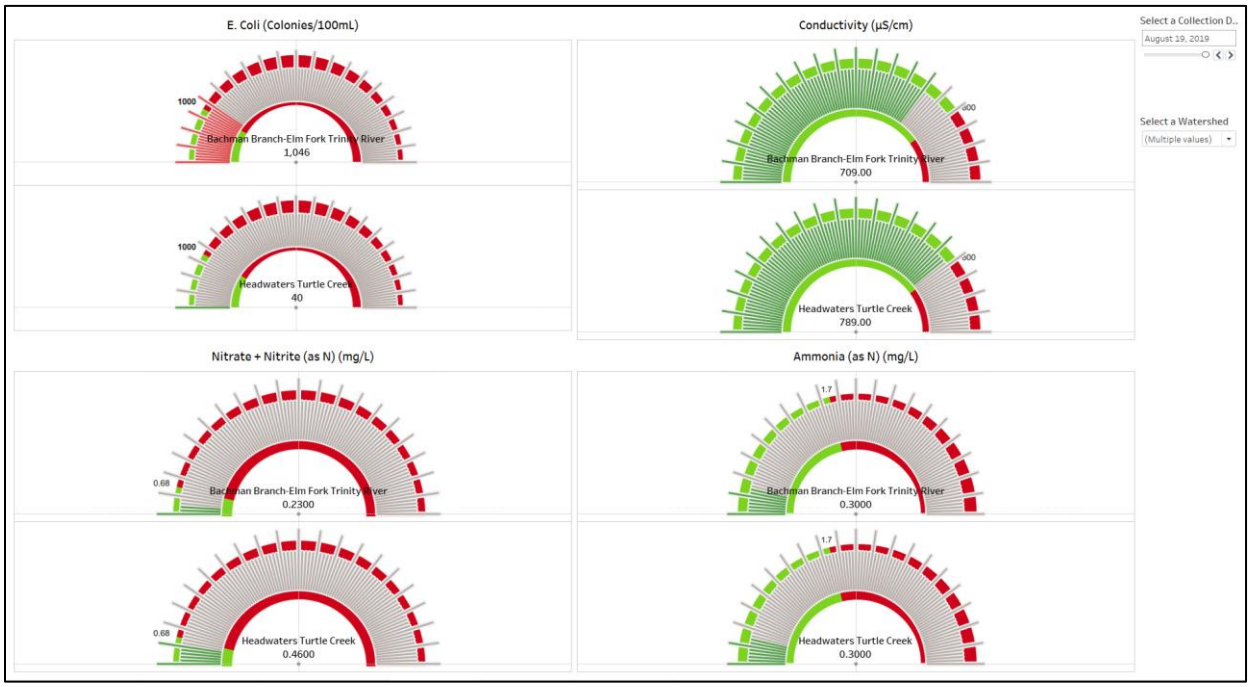
WATERSHED COMPARISON

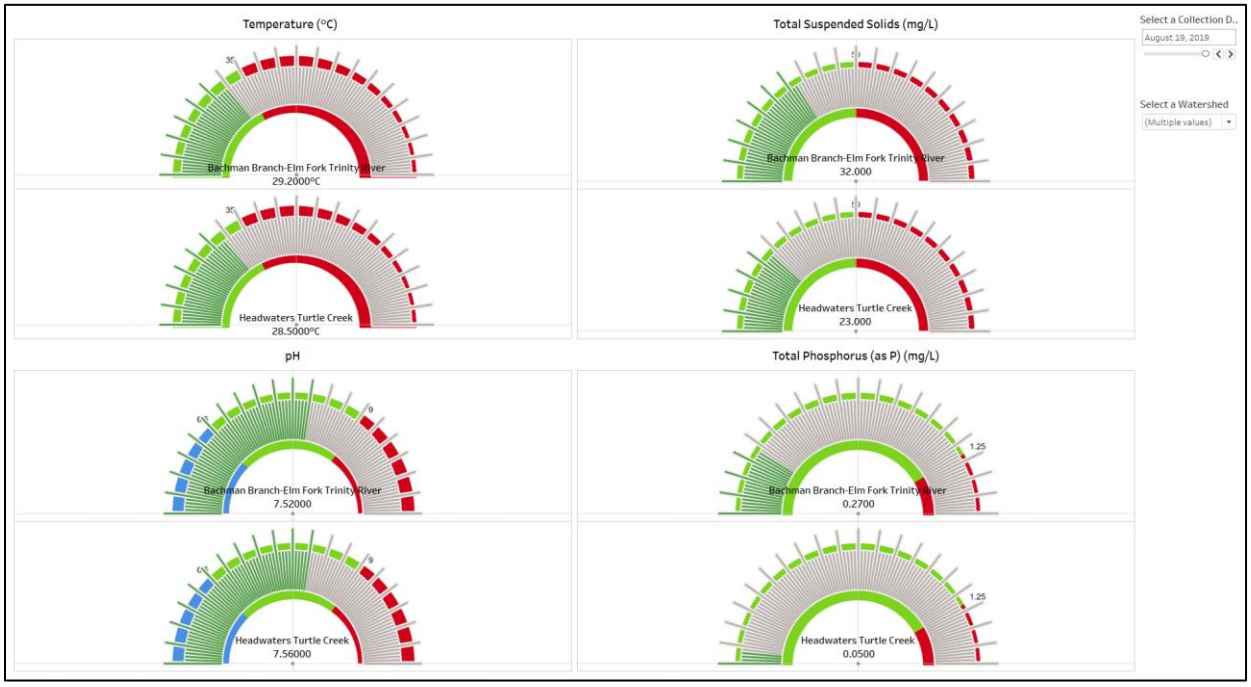


Business Insights:

- Spikes in E. Coli are seen in 2012 in both branches as well as 2014 in Headwaters Turtle Creek.
- The Bachman Branch watershed saw a spike in Ammonia and Phosphorus in May of 2015.
- In August 2011, the level of Total Suspended Solids rose well above the parameter level in Headwaters Turtle Creek.
- Nitrate + Nitrite levels are above the acceptable parameter for both watersheds in almost all years, but were especially high in March and April of 2014 and 2015, and April and May of 2018 for the Bachman Branch.
- Conductivity levels are above the acceptable parameter goal for both watersheds in all years.
- High temperatures are not much of a concern as all water samples for the past decade have temperature levels within the acceptable parameter.
- When looking at the composite score, it seems that Bachman Branch scores higher than Headwaters Turtle Creek for most years. As each of the pollutants has equal weight in the computation of the composite score, there must be some pollutants in the Bachman Branch that perform better than in Headwaters Turtle Creek. Most of the pollutant distributions look the same, but after excluding extreme values of E. Coli and Nitrates + Nitrites from the time series, we can tell from the chart that most of the values for Bachman Branch are distributed around the reference line while Headwaters Turtle Creek values are more decentralized. The same thing appears in the Nitrate and Nitrite. That might be the reason why the score for Bachman Branch is better.
- Box and whisker charts are computed for each watershed. The grey area is calculated as 1.5 times the IQR which means the smaller the grey area, the more condensed the data is. Even with the effects of outliers, it is obvious that most of the pollutants performs better in Bachman Branch. This indicates that pollutants in Bachman Branch are more stable.

GAUGE CHARTS 1 AND 2





The gauges are split into two dashboards to make them easier to read and interpret, but the set-up of the dashboards is identical otherwise.

Business Insights:

- The gauges allow the viewer to see where pollutant values for each water sample stand in comparison to their minimum and maximum acceptable parameters. As discussed before, the only pollutants that are performing outside of their goal levels in the most recent water sample are E. Coli and Conductivity. However, after changing the water sample date on the "Select a Collection Date" filter, the values change and show many more readings that are outside their parameter goals.