Data Analysis and Visualization

Contest Statement:

Common techniques in *data analytics and visualization* include data acquisition, data cleaning, handling missing values, wrangling, data integration, simple data analysis (e.g., outlier detection, identifying collinearity between variables, dimension reduction, summary statistics, and determining skewed attributes), statistical model building, finding patterns, finding clusters, plotting geo-temporal points on a map, data visualization, visual analytics, problem solving, making recommendations, and providing quantitative analysis to support decision making.

Contest - Case Study – Domain Questions to Answer

This section contains background information on the domain of *marine ecology* which is necessary to complete the contest's tasks. Review the following background material on hypoxic environments. Note: this is currently an open, real-world challenge, not a solved problem. You can expect existing datasets to be less than ideal, dirty, contain missing values, misaligned with current real-world situation, etc. **Turn in answers to these five case study questions as part of the contest (10% points possible).**

Review background domain information:

- https://gulfhypoxia.net/about-hypoxia
- Optional URL http://coastal.ohiodnr.gov/portals/coastal/pdfs/owc/tech/owc_techbull3_Hypoxia.pdf

  Question 1: Define hypoxia. What is the dissolved oxygen concentration threshold that identifies hypoxia? _____

  Question 2: Which species are affected? How are these species affected? _____

Review the animation of hypoxia forecast data from Lake Erie's prior seasons:

- https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/hypoxiaWarningSystem.html
- https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/archive/dissolvedoxygen2018.gif

  Question 3: Where and when is hypoxia likely to occur in Lake Erie? Describe the general locations or provide an annotated map. _____

Remote monitoring:

- https://www.ndbc.noaa.gov
  Question 4: Are current NOAA buoys/stations and sensors located in appropriate locations for monitoring and surveillance Lake Erie's hypoxia? If so, which station IDs are in pertinent locations? _____

- https://www.ndbc.noaa.gov/ship_obs.php
  Are there shipping lanes through the area that might provide sporadically-sampled monitoring data? Skip this question for now due to the current lake conditions.

Historic data 2004-2007:

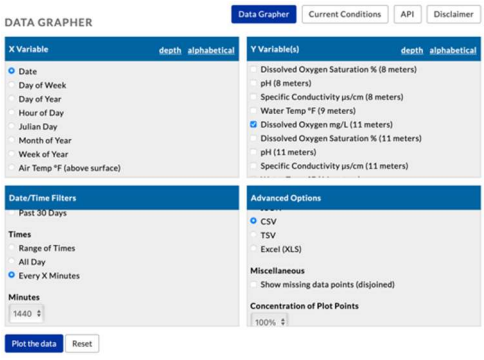
- https://www.glerl.noaa.gov/res/projects/ifyle
- https://www.glerl.noaa.gov/res/projects/ifyle/data.html
- https://www.glerl.noaa.gov/res/projects/ifyle/data/data.mooring.html
- https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/data.ysi.html
- https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/ysi/eriemap.html
  Review these details on the GLERL collaboration which has attempted to capture and monitor datasets related to hypoxia from the last 15 years in Lake Erie. Select a buoy location that appears to experience hypoxia events throughout the year (e.g., location Y18). **Review the metadata** and challenges of collecting this type of data.

- https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/ysi/ysi_metadata.txt
  Question 5: These datasets contain very useful information (sensor depth, dissolved oxygen concentrations, temperature, turbidity, and chlorophyll) but have not yet been cleaned. Describe the errors that are known to exist within the dataset. _____

Contest – Tasks

Select and complete as many of the following tasks as you can. Each task is equally weighted, and teams earn points for each successfully completed task. Each task will require a dataset to be extracted, manipulated, cleaned, and visualized. The exact approach, software, libraries, and solution is left up to the students, in keeping with the spirit of the competition. Your team must perform your own analysis for the exercise and cannot reuse solutions or code produced by others. You are welcome to consult online tutorials and reference guides regarding your software if you wish.

Task A



Hypoxia is most likely to occur when a lake is stratified into distinct layers (e.g., the top and bottom layers of the lake are different temperatures and the water does not mix). In turn, the oxygen on the bottom of the lake eventually reduces to less than 2 mg/L and results in the death of aquatic organisms at the bottom of the lake. Use GVSU's Muskegon Lake buoy datasets to identify all of the hypoxia events in Muskegon Lake from 2011 to 2022. Clean and filter the data if needed, then use software or programming to determine all of the days in that year that have a dissolved oxygen reading less than 2.0 mg/L at the bottom of the lake. Aggregate and visualize the total number days with hypoxia conditions for each year (e.g., ~54 days in 2011).

Dataset URL: https://www.gvsu.edu/wri/buoy/data-index.htm

Hint: Check *Date* as the X variable, *Dissolved Oxygen mg/L at 11m* as Y, *All Dates* every *1440 minutes* as the time range (one observation at midnight per day), and export the data as a *CSV*.

Deliverable: Create a chart that compares the total number of hypoxia days for each year. Also, write a sentence stating which year had the most days with hypoxia.

Task B

A harmful agal bloom (HAB) is identified when there is a spike in Phycocyanin and Chlorophyl levels near the surface of a lake. In practice, finding 'spikes' in these readings are relative to each time the sensors are recalibrated. Some years, the sensors are calibrated to different ranges. For simplicity, assume a spike in Phycocyanin is any reading greater than 20.0 Cells/mL (at 2 meters) and Chlorophyl at readings greater than 10,000 µg/L (at 2 meters). Use GVSU's Muskegon Lake buoy dataset to identify all of the possible HAB events between 2011 and 2022. Clean and filter the data, then use software/programming of your choice to identify and list all dates with both of these conditions.

Dataset URL: https://www.gvsu.edu/wri/buoy/data-index.htm

Hint: Check *Date* as the X variable, *Chlorophyll µg/L (2 meters)* and *Phycocyanin Cells/mL (2 meters)* as Y, *All Dates* every *1440 minutes* as the time range (one observation at midnight per day), and export the data as a *CSV*.

Follow-up question: do HAB events appear to be 1) short in nature and spread arbitrarily throughout the last decade or 2) clustered into a few rare periods with many HAB days in a row?

Deliverable: Create a chart that proves **which year** had the most HAB days. Provide a sentence answer to the follow-up question.

Task C

Analyze one year of hypoxia events in Lake Erie based on a high-dimension dataset. Create a list or visualization that details the specific days and time windows that the dissolved oxygen fell below 2 mg/L back in 2007. Analyze the data only for the location at monitoring station Y18 covering the data from that whole year. Consider using Tableau, Python, or Excel to generate the list of hypoxia readings. Note: a date of `205.5208` indicates the 205th day of the year (July 24) at 12:30pm. Can a DO reading be negative? If not, you may have some data cleaning to do with the negative values.

Dataset URL: https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/ysi/2007/Y18.txt

Metadata URL: https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/ysi/ysi_metadata.txt

Deliverable: Provide a few sentences presenting the total number of hypoxia readings you found and also the data cleaning you did to the dataset, if any.

Task D

Visualize lake conditions back in 2007 on Lake Erie based on datasets from multiple buoys/stations (Y07, Y08, Y09, Y10, Y11, Y13, Y17, and Y18). Create a geo-spatial map that uses *color* and *size* to display the lake's physical properties for all of these locations at a single point in time. Select two attributes (water temperature, DO, turbidity (water clarity), or chlorophyll), and generate a chart

that displays both of the readings from all eight buoys at a similar timestamp. Generate one chart (for all eight locations) at one timestamp of your choice. Note: observation timestamps should be approximately the same for each station. Include the timestamp you used in your chart title/caption. Consider using a python mapping library or Tableau.

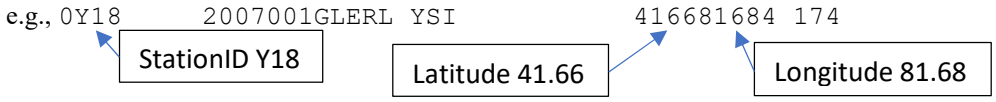
Dataset URLs:

https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/ysi/eriemap.html

https://www.glerl.noaa.gov/res/projects/ifyle/data/Mooring/ysi/Station.php?sta=All&year=2007

Geospatial coordinates: https://www.glerl.noaa.gov/res/projects/ifyle/data/stations.lmd

Geospatial metadata: https://www.glerl.noaa.gov/res/projects/ifyle/data/stations.lmd.fmt



Deliverable: Create a map-based visualization that shows any two attributes of your choice for all eight stations at one timestamp of your choice.

Task E

Visualize potentially hazardous conditions for current (live) locations of ships that are reporting scientific observations to the National Data Buoy Center (NDBC). Fog forms when the air temperature is lowered to within 5°F of the dew point. E.g., 68°F air temperature and 66°F dew point indicates possible fog. Wave heights greater than 7.0' and wind speeds greater than 20 knots can also create challenging water conditions. Create a geo-spatial map that plots the coordinates of each ship as well an indicator if the ship is potentially experiencing any of the three difficulties (possible fog, high wave height, or high wind speeds). E.g., use a red versus black points on the map to indicate which ships are facing hazardous conditions. Generate one chart/map showing all active ships (e.g., during a recent one-hour window on today's date). Consider using a python mapping library or Tableau maps.

Dataset URL: https://www.ndbc.noaa.gov/ship_obs.php

Deliverable: Create a map-based visualization that showing the location of all current ships and markup each ship if it is experiencing any of the hazardous conditions.

Contest - Deliverables

Solve the tasks in an individual or two-person team – sharing completed solutions with other teams is not allowed. You are welcome to either utilize existing software to manipulate/analyze data, write your own code, or follow any other analytical methodology you prefer to arrive at a solution to each task. To complete this challenge, create and turn in a document that answers all the background questions from part *Contest - Case Study* as well as many tasks as you can from part *Contest - Tasks*. Also, be prepared to turn in or demonstrate any other supporting files such as computing code (e.g., Python scripts), Tableau workbooks (.twbx file), intermediate datasets (e.g., CSV files), etc. Do not upload any part of this contest or your deliverables to online locations (e.g., GitHub).

Contest - Software recommendations

If you choose to use **Python** for this contest, SciPy is a set of widely used packages for managing, analyzing, and visualization large-scale content. It consists of libraries for data structures such as DataFrames and analysis (pandas), N-D arrays (NumPy), 2D graph plotting (Matplotlib), scientific analysis (SciPy), etc. In addition to matplotlib, other popular python libraries include *ggplot (any R fans in the group), plotly, seaborn, pygal, bokeh, geoplotlib, etc.* These packages differ by their customization, expected data input structures, charts available, export formats (e.g., svg), interactivity, dependencies (web integrated), etc. See:

https://www.python.org/about/gettingstarted
https://docs.python.org/3/tutorial
https://scipy.org
https://matplotlib.org

Tableau and **PowerBI** are leading software tools for visual analytics and rapidly create interactive visualizations. You may download Tableau Desktop from tableau.com with a 14-day free trial, or perhaps Tableau Cloud. Students (worldwide) get free renewable year-long licenses for Tableau Desktop (with a valid university email address), see https://www.tableau.com/academic/students. Getting the license key will likely only take a few seconds. If you require brief tutorials on how to use the GUI and its functionality, see the APPENDIX below, or visit https://www.tableau.com/learn/training. Do NOT post and make your work visible for others to see within the cloud-based Tableau Public. Tableau connects with a variety of underlying dataset formats: offline data file on your hard drive, online databases, an online data server, or the default practice datasets (e.g., try out the World Indicators dataset). Note how the columns from the dataset are available on the left pane, you can drag columns to the middle pane to set the x-position y-position color size (which automatically updates the chart graphic), dragging columns to the filter pane generates dynamic filters within the visualization, and the right pane allows you to change the chart type depending on the columns already selected. Users explore their dataset by creating a sheet and pairing various combinations of dimensions and measures. For each combination of columns that you explore, switch the visualization to several options (including Tableau's recommended chart choice which is highlighted with a red box on the list of possible charts on the right pane). You might use this contest as an opportunity to become familiar with Tableau, how to connect to a dataset, how to create charts (sheets), and how to create interactive dashboards.

The bulk of the work for each task can be done in **Excel**. Alternative analytics or visualization software you may find useful include Google charts, Excel, R, SAS, SPSS, Weka, and/or Power BI. Students are encouraged to use any software and online resources they wish, including Google, to learn skills needed to complete the challenges.

Contest - Rubric

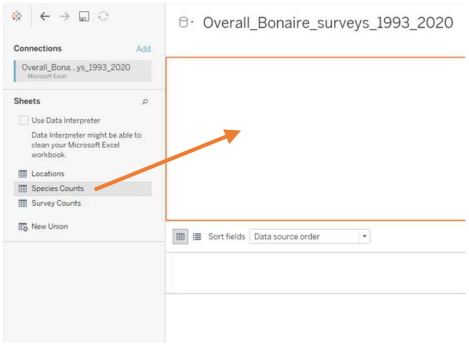
The following weighted rubric details how the results will be evaluated. The final submission should be organized in a single Word or PDF file (with legible screenshots if needed). Provide solutions that include all the required deliverables for each task that you are able to complete. Partial solutions will be given up to half credit. Sort your tasks A-E in their proper order. Clearly indicate which tasks you attempted and which tasks you skipped.

1. Background domain knowledge from *Contest - Case Study* (10%)
2. Task A through Task E (18% for each task solved)

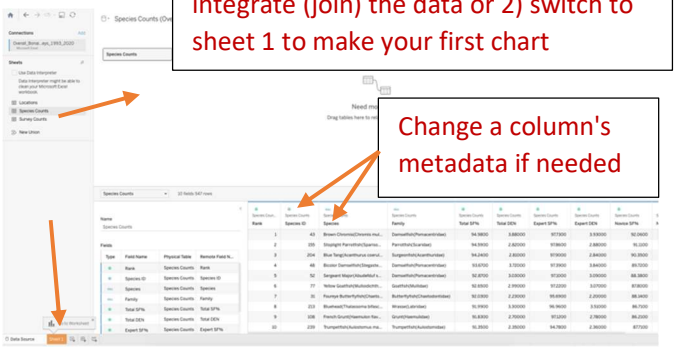
Appendix: Getting Started with the Tableau Desktop Interface - A Quick Guide

Select a . csv or Excel

Load one spreadsheet tab from a data source at a time. Dragging two tabs will force Tableau to perform an inner-join to integrate the data.

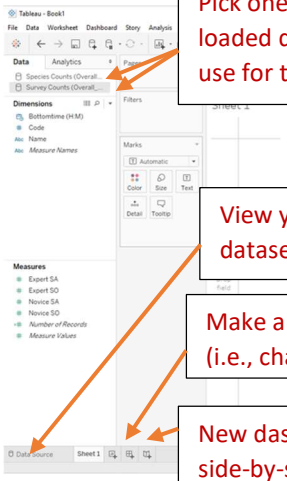


1) add a second tab from the file to integrate (join) the data or 2) switch to sheet 1 to make your first chart



Change a column's metadata if needed

Pick one of your loaded datasets to use for this chart

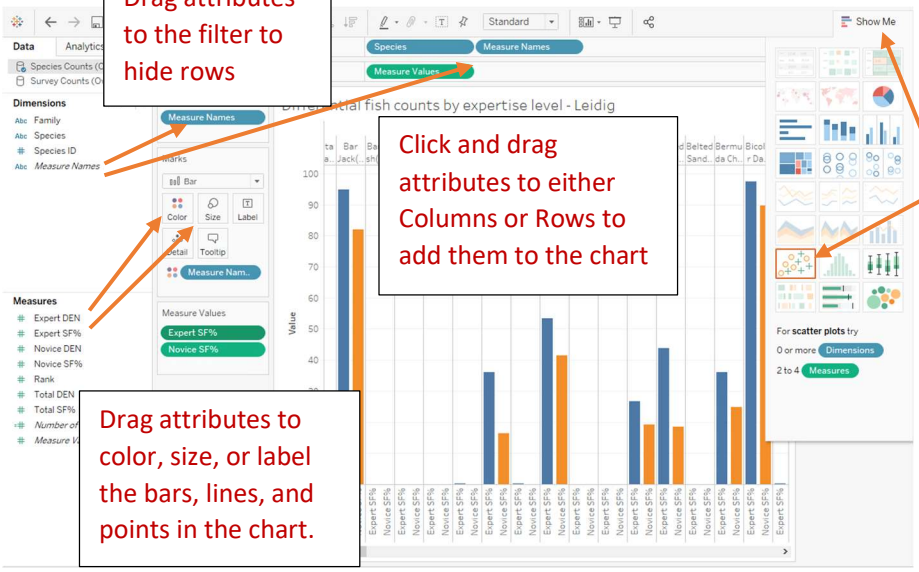


View your loaded datasets

Make a new sheet (i.e., chart)

New dashboard (2+ side-by-side charts)

Drag attributes to the filter to hide rows



Click and drag attributes to either Columns or Rows to add them to the chart

Choose the type of chart you want to generate. Based on your data, Tableau will recommend a chart type with the red outline in the "Show Me" window.

Drag attributes to color, size, or label the bars, lines, and points in the chart.