Contest: Data Analysis and Visualization

PLEASE DO NOT OPEN UNTIL INSTRUCTED TO DO SO

IMPORTANT:

Any solution received that includes any **IDENTIFYING** information will not be eligible for anything beyond "Honorable Mention."
Examples include any of the following:

- Names of the individuals or school names in the solution
- Names of the individuals or school in the "properties" of the file
    - Each program, Adobe, Microsoft, etc. contains metadata about the document. Remove ALL of these from your document. If you do not know how please ask someone!

Any solution that does not provide accurate citing of professional resources will be removed from consideration.

Examples include any of the following:

- Copying and pasting diagrams and images from a website
- Using descriptions and product data verbatim from source
- When in doubt, cite your source

Data Analysis and Visualization Contest

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.

Each team may consist of one or two individuals and must have at least one computer to complete your project. The contest is written from the perspective of guiding someone who is new to using Tableau. Appendix A contains a quick-start guide to Tableau's interface. You are welcome to use other software and skip the Tableau focused tutorial instructions found in Appendix B – you are only required to turn in your answers and a visualization (using any software of your choice) for the required tasks.

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 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 as many tasks as you can. Also, turn in any other supporting files such as computing code (e.g., Python scripts), Tableau workbooks (.twbx file), intermediate datasets (e.g., CSV files), etc. to showcase your work. Do not upload any part of this contest or your deliverables to online locations (e.g., GitHub).

Resources - Software recommendations

It is recommended to download Tableau Desktop (free year-long license) or else use online Tableau Cloud for the contest (see: https://www.tableau.com/academic/students). Alternative analytics or visualization software that students may find useful include Excel, R, Python, SAS, SPSS, Weka, and/or Power BI. Students are encouraged to use any software and online resources they wish, including Google, to learn and complete the challenges.

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 generate 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 A** below, or visit https://www.tableau.com/learn/training. Do NOT post and make your work visible for others to see. 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. Some of the work for each task can be done in **Excel**.

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 solved tasks A-G in their proper order. Clearly indicate which tasks you attempted and which tasks you skipped. The earliest timestamp for the final pdf answer submission will be used to break any ties.

Task A through Task G (equally weighted for each task solved correctly)

Contest – Case Study Background

Bonaire is an island in the Caribbean Netherlands (next to Aruba and Curacao) located a few miles off the coast of Venezula. It is well-known for its healthy coral reefs, salt industry, natural park, and tourism based on SCUBA diving and water sports. There are over 100 coral reefs locationed around the coast of Bonaire. Let's use visually explore marine biology data from this island (provided by www.reef.org). Over the past several decades, the REEF organization collected underwater surveys from SCUBA divers at each of these locations. The surveyors were marine biology experts as well as novice citizens scientists.

In the dataset we will use, **Name** refers to a geographic area of coral reef that was surveyed. The corresponding **Code** is a unique number that was generated for each reef location. Recall from your basic biology courses that the Taxa for family and genus are broken down into individual species. On these survey reports, **SA** columns indicate that the surveyor reported the ***S****pecies and the species' **A**bundance level*, while **SO** columns indicate that the surveyor *reported the **S**pecies **O**nly* and not the abundance level of the species. The sighting frequency (**SF%)** refers to the percent of reports in which surveyors reported that species. E.g., experts report seeing the Blue Tang species on 97.9% of surveys. **DEN** refers to the density (abundance) of the species when it is observed. E.g., 1 means the species was solitary and only one was observed, while 4 means a very large school of this fish species was observed.

Given the Overall_bonaire_surveys_1993_2020.xlsx file, solve as many of the following tasks as you can. Create visualizations (e.g., a sheet within Tableau) that solve the following tasks. Turn in a Word or

PDF file with a **written answer** to each task and a **screenshot** of your visualization that visually supports your answer. The optional Appendix B provides a very helpful walkthrough tutorial for using Tableau to load the data, understand the case study, and answer these types of questions. If you are new to Tableau, I highly recommend you begin by reading the **Context Tasks** section, and then proceed through Appendix B, which will prepare you to solve the contest's actual tasks.

Contest – Tasks to solve

Task A: Novice and expert surveyors are able to identify different sets of fish, which results in differing reports. What is the top **species** that novices over-reported most frequently compared to experts? I.e., in comparing the SF% for species that are more often reported by novices, which species has the largest difference between the two groups?

Task B: Which **species** is ranked last with the highest Avg(Rank) (i.e., least likely to actually be identified and reported)?

Task C: What are the top most likely **family** to be observed - with the lowest Avg(Rank)?

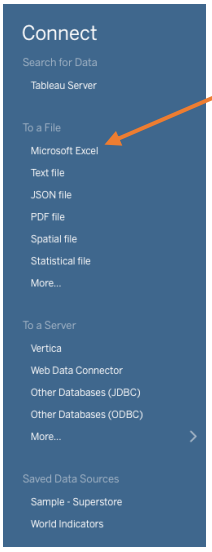
Task D: One hotel (cleverly) paid a group of expert marine biologists to survey the reef closest to their hotel. That reef is now considered one of the most biodiverse reefs in the Caribbean due to scientists' expertise and the hotel's ploy. What **reef name** had the largest number of total expert surveys completed (Expert SA and Expert SO surveys)? How many total expert surveys were completed?

Task E: What **reef name** had more novice surveys completed at that location compared to expert surveys? How many **more** novices completed surveys at that location than experts?

Task F: What is the northernmost location (i.e., **name** of the location)?

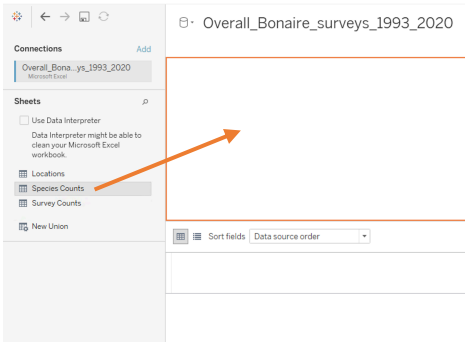
Task G: Based on this dataset and your general background knowledge on island life, why might the Southwest coast of the island be heavily surveyed and not the East coast? Could this bias the type and abundance of species within the surveys of this trusted dataset?

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

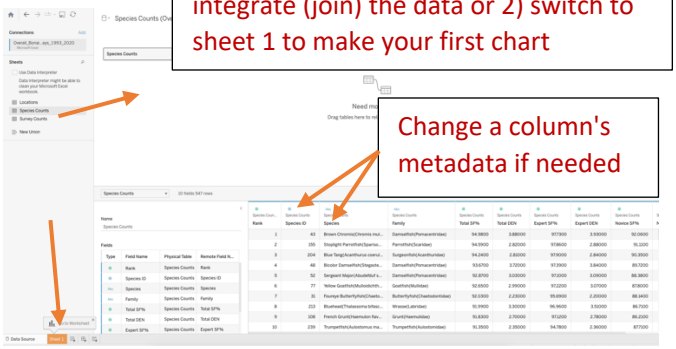


Select a .csv or Excel

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.

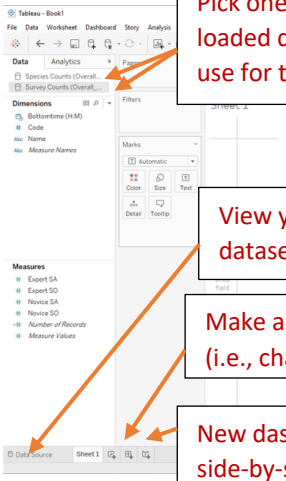


1) add a second tab from the file to integrate (join) the data or 2) switch to sheet 1 to make your first chart



Change a column's metadata if needed

Pick one of your loaded datasets to use for this chart

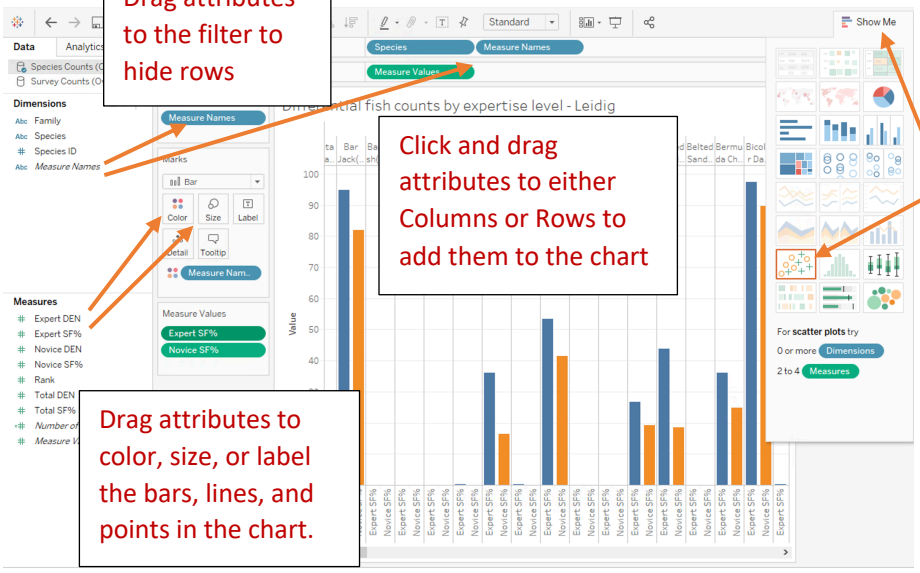


View your loaded datasets

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.

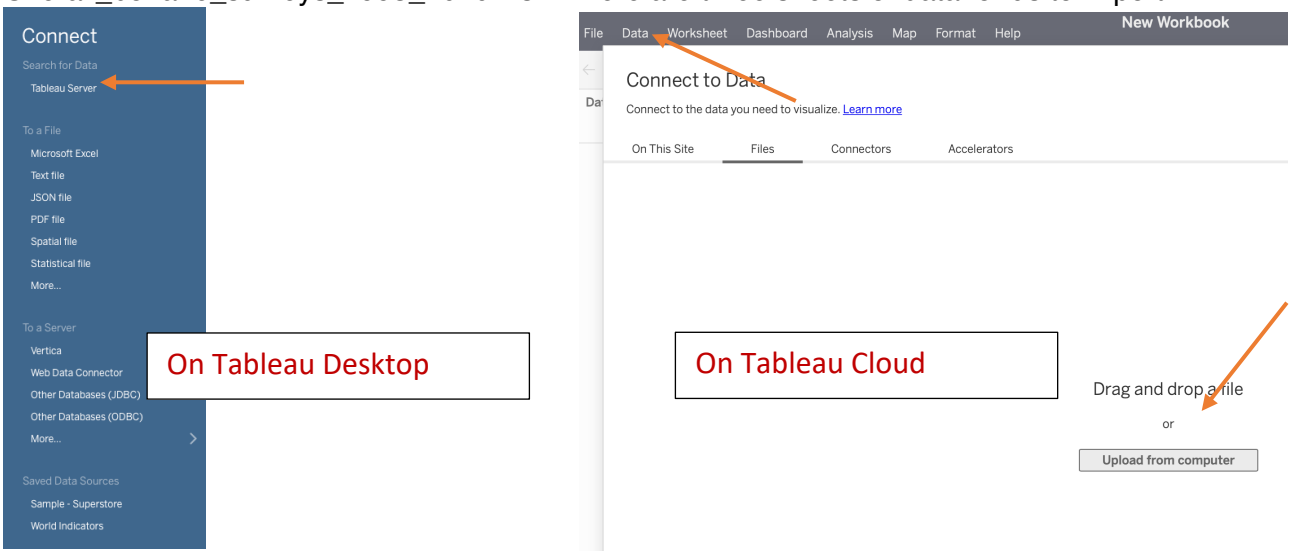
Drag attributes to color, size, or label the bars, lines, and points in the chart.

Appendix B: Optional Walkthrough for Solving These Tasks in Tableau

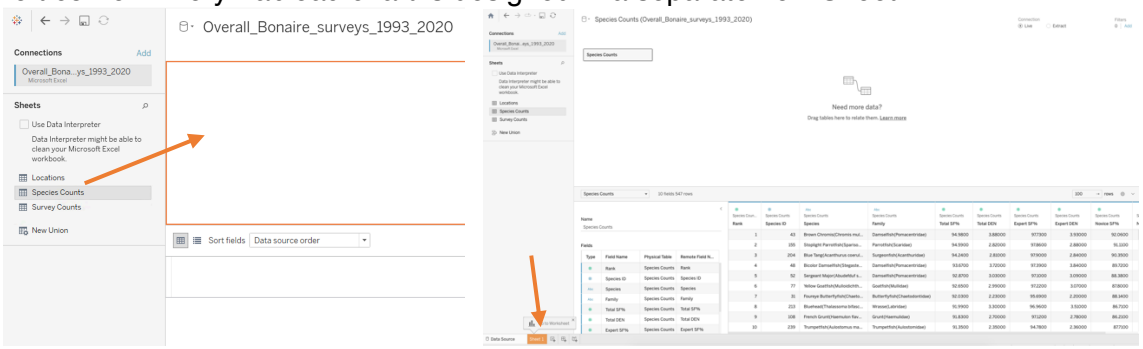
NOTE: This is a ***teaching guide*** to increase your familiarity with Tableau. Don't turn in the results from Appendix B! Your actual tasks to complete for the contest are in the **Contest – Tasks to solve** section.

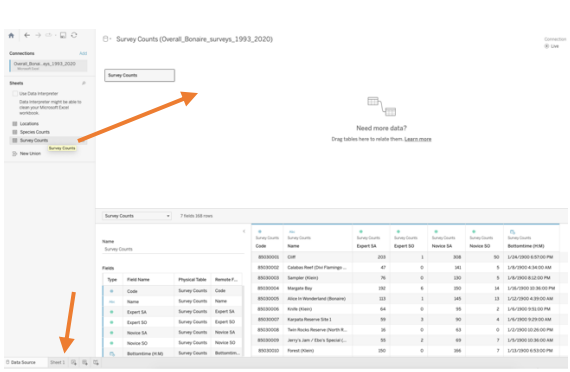
Bonaire is an island in the Caribbean Netherlands (next to Aruba and Curacao) located a few miles off the coast of Venezula. It is well-known for its healthy coral reefs, salt industry, natural park, and tourism based on SCUBA diving and water sports. Let's use Tableau to explore Bonaire's Marine Biology data (from www.reef.org).

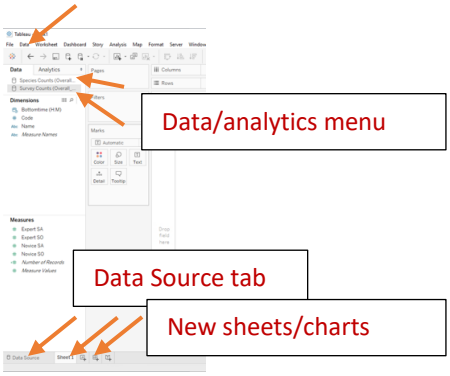
1. Download the Overall_bonaire_surveys_1993_2020.xlsx file to your computer (email leidijon@gvsu.edu to receive a copy of the dataset if you do not have access to the files). You can open the file in Excel to check the data.
2. Import the file in Tableau - a collection of fish surveys reported by experts and novices. Use the Connect panel on the left to connect to a Microsoft Excel file, find and open your downloaded file: Overall_bonaire_surveys_1993_2020.xlsx. There are three sheets of data for us to import.



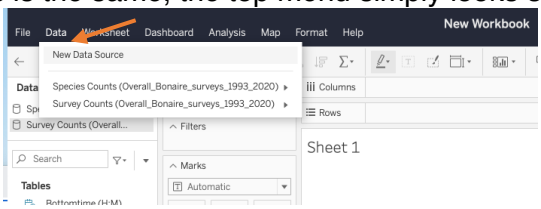
3. We will start by adding two sheets of data, one at a time. First, select the *Species Counts* Excel worksheet and drag it to the pane on the right. Then click on Sheet 1 on the bottom menu. A *sheet* is the Tableau word for a single chart. Under the Data menu at the top of your screen, select a New Data Source, find your excel file a second time, and then add the other *Survey Counts* worksheet by dragging it to the right pane. Go back to the Sheet 1 chart. After this step, both sheets (*Species Counts* and *Survey Counts*) should be visible in the top left Data/Analytics menu. Ignore the *Locations* datasheet from the Excel file for now. The data can always be reviewed in the Data Source tab on the bottom-left corner. Every Tableau chart is designed in a separate new Sheet.







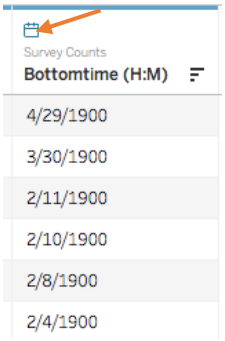
On Tableau Cloud, the process is the same, the top menu simply looks slightly different.

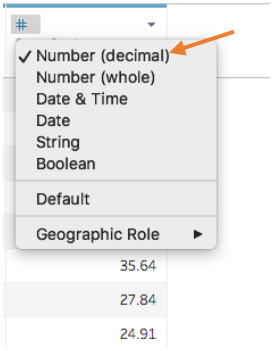


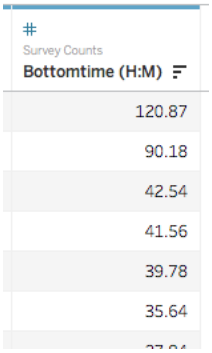
Answer these questions to yourself but DO NOT turn in:

a. How many rows and columns does the Species Counts Data Source contain? Hint: use the Data Source tab.
b. How many rows and columns does the Survey Counts Data Source contain? Hint: use the Data Source tab.

When importing the Survey Counts sheet from Excel, note that the variable Bottomtime is a number of hours and minutes (which Tableau sometimes mistakes for a date). In the data source pane, let's switch that back to Number(decimal).





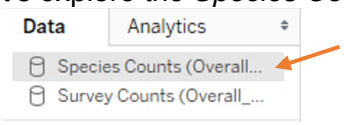


c. For more details on the data source, checkout: https://www.reef.org/db/reports/geo?end_date=2020-02-10&format_type=chart&group_type=species&language=common&region_code=TWA&start_date=1993-01-01&zone_map=0&zones=8503

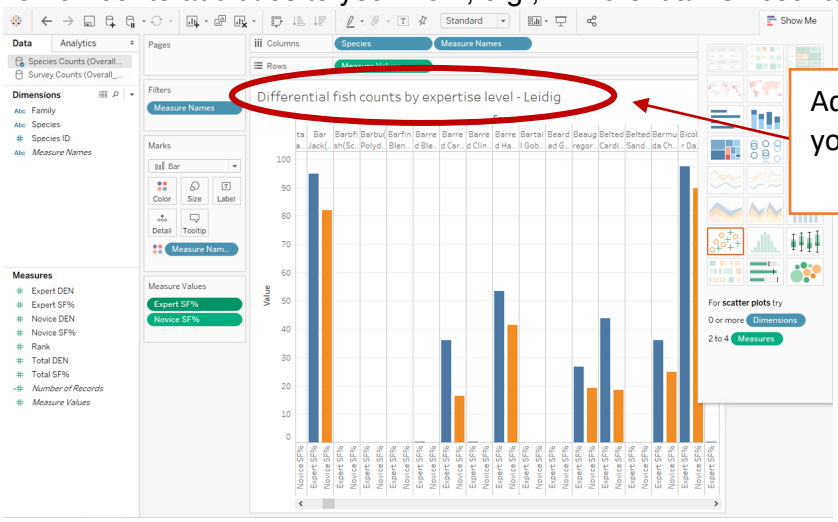
4. Use *Sheet 1*. The columns from the dataset are available on the left pane, you can drag columns to the middle pane to set the x-position/y-position/color/size (which automatically updates the interactive graphic), dragging columns to the filter pane generates dynamic filters within the visualization (allowing you to filter out undesired data), and the right pane allows you to change the chart type depending on the columns already selected.
5. First, explore your 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 visualization which is highlighted by a red square on the list of possible charts on the right pane depending on the columns you selected).

Next, use Tableau to explore the dataset. I recommend making a new visualization for each task, that way you can go back to earlier steps.

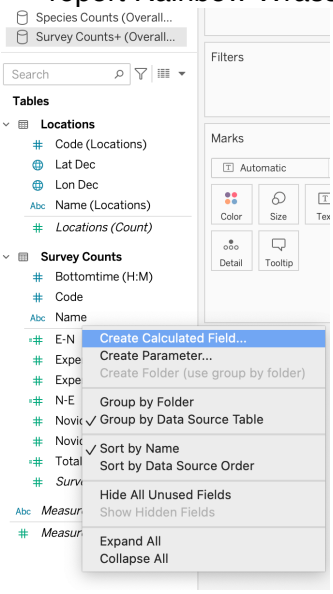
First, let's explore the *Species Counts* data. In sheet 1, select the Species Counts dataset you have imported.

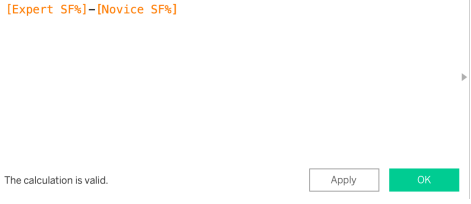


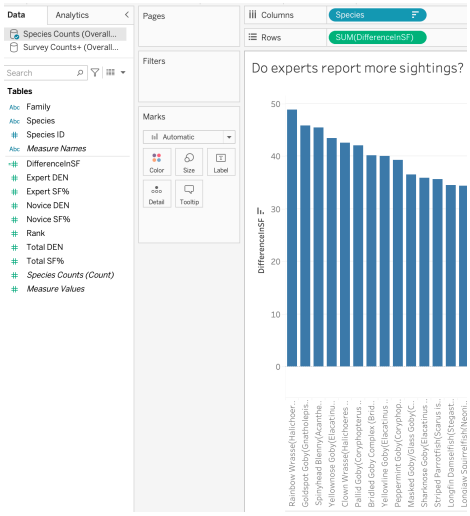
6. Compare *novice sighting frequency* to *expert sighting frequency* for each *species*. Find *5-10 species* that experts seem to identify more frequently than novices. Hint: try a side-by-side bar chart and show individual datapoints (click Analysis in the top menu and then unselect Aggregate Measures).
    d. You would ideally represent the insight in the simplest visualization that gets the story (key takeaway) across. Write a short title that clarifies what the key takeaway is for the reader. Remember to add titles to your work, e.g., "Differential fish counts by expertise level - Leidig".



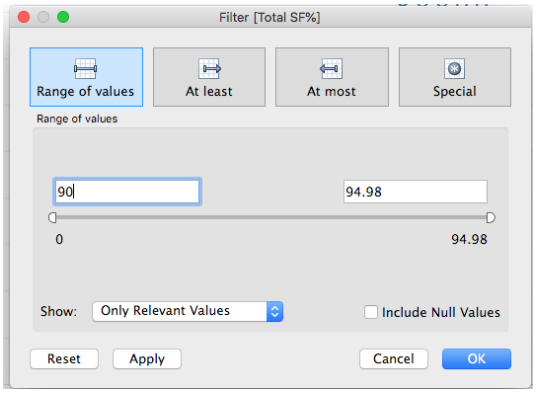
7. Using the new chart, how much more do *experts* seem to identify these species more *frequently* than *novices*? A calculated field would make this much easier to solve. Right click in the left pane within the Tables section and select *Add Calculated Field…* Write a simple calculation to generate a new column (*DifferenceInSF*) for each row in the data table that subtracts Novice SF% from Expert SF%, calculating how much better experts are at sighting each species than a novice. It looks like experts identify and report Rainbow Wrasse more often than novices.





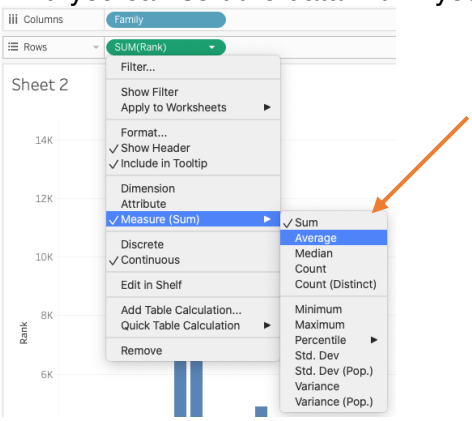


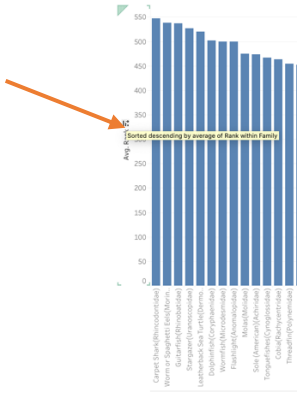
8. Create a new sheet (Sheet 2). I recommend making new sheet and chart for each question. Some species are very prevalent and are observed on almost every fish survey. Which species are seen on at least 90% (or more) of total surveys? Show all of the ~11 species. Hint: create a chart for *species* and *TotalSF%*, drag the field *Total SF%* to the Filters pane, and then only show species found at least 90% of the time.



9. Recall from Biology, multiple species are organized into a family of similar organisms (Homo sapiens fall under the Hominidae family). E.g., *Brown Chromis* and *Bicolor Damselfish* are both grouped into the *Damselfish* family. The *species* with a *rank* of 1 is seen most frequently and the *species* with a *rank* of 547 is seen the least frequently.
    e. Generate a **new sheet**/chart that shows the *average* (avg) *rank* of the high-level *family* groups.

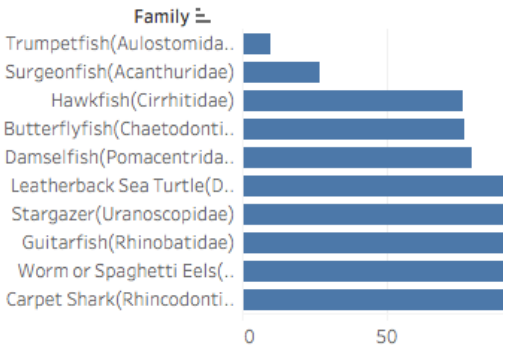
    Hint: you can sort the data within your charts to make it easier to answer the questions.



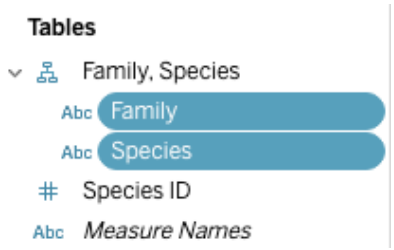


    f. Hint: generate a graph of *family* compared to Avg(*Rank*). Tableau defaults to calculating the Measure (Sum) instead of Avg – but that can be changed in the dropdown from *SUM(Rank)*.

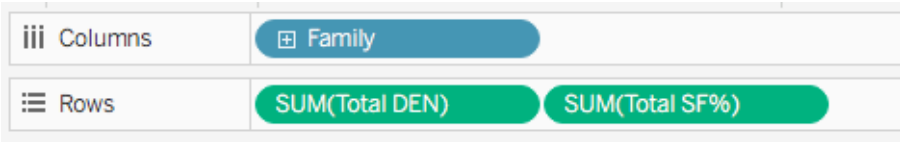
    g. What are the top 5 most likely families to be observed - with the lowest Avg(Rank)?

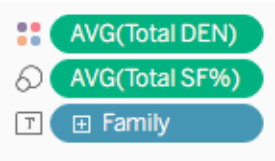


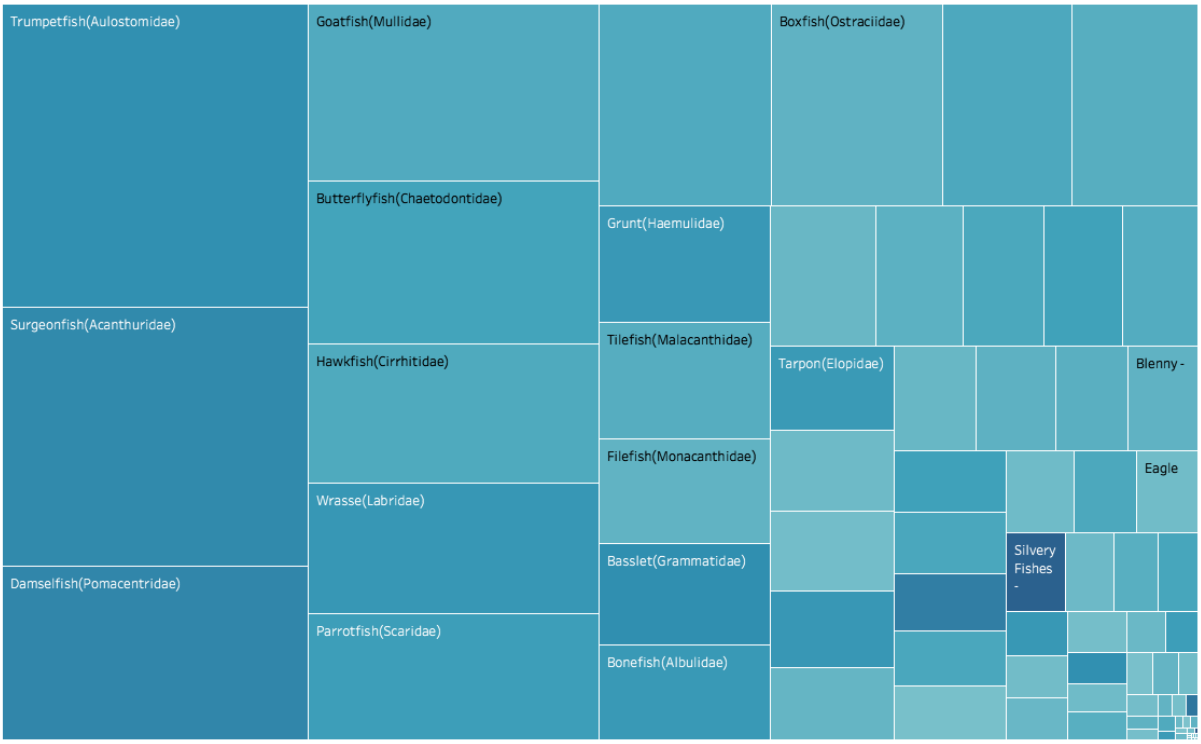
10. Dragging the *Species* dimension on top of the *Family* dimension in the left pane will create a hierarchy. Thus, the *Family* category will be the parent to the *Species* sub-category (Family,Species).



h. Generate a new Treemap sheet showing the average *total sighting frequency* and average *total density* for each *family,species* using size & color. It may be helpful to set *SF* to area and *DEN* to color.

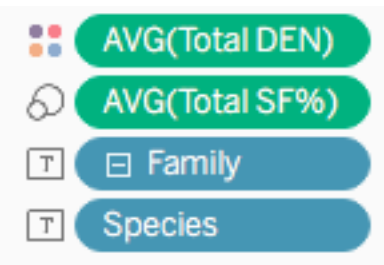


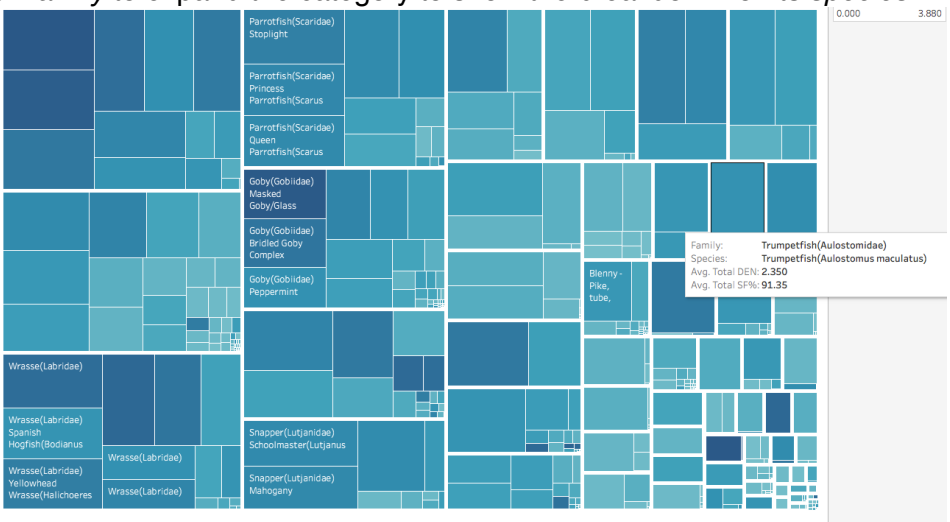




i. Which *families* have the highest *densities*? A schooling family like *silvery bait fishes* perhaps?

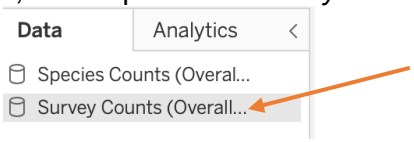
11. From the same style chart, click *family* to expand the category to show the breakdown for its *species*.



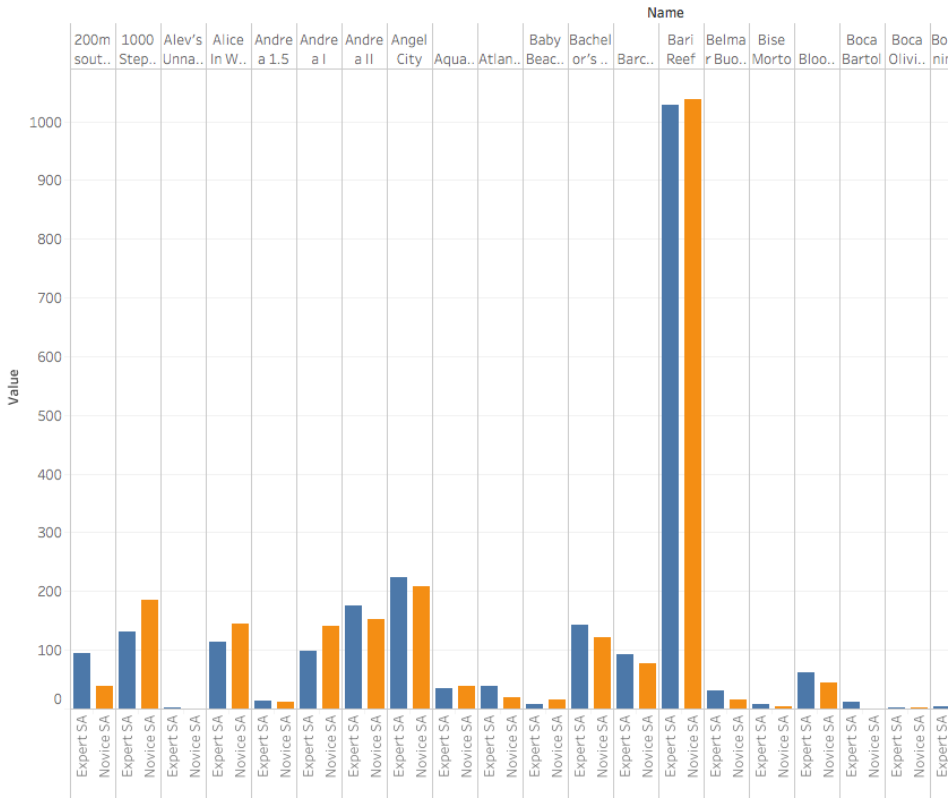


j. Find the Trumpetfish sub-category – it may take some hunting! Let's compare the Trumpetfish *family* to the Damselfish *family*. There is only one *species* of Trumpetfish in our data while there are 14 Damselfish *species*. Pretty neat!

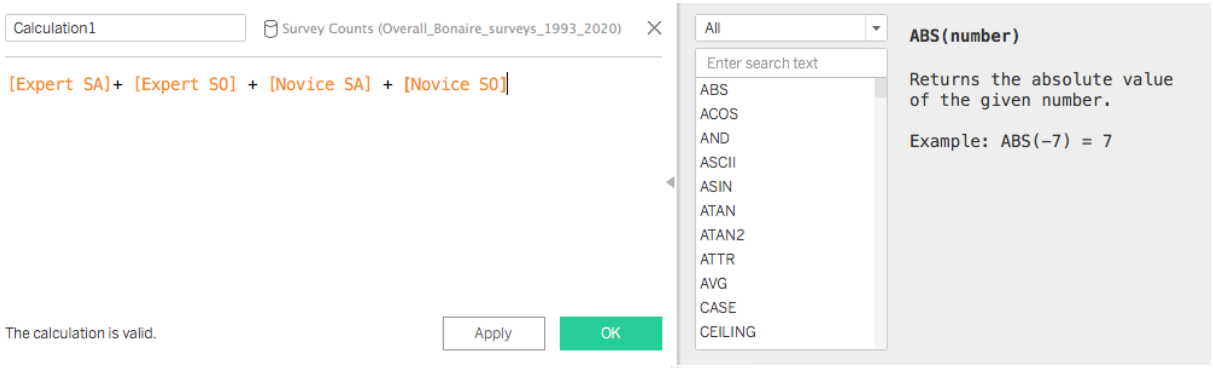
Next, let's explore the *Survey Counts* dataset. Create a new sheet and select the Survey Counts dataset.

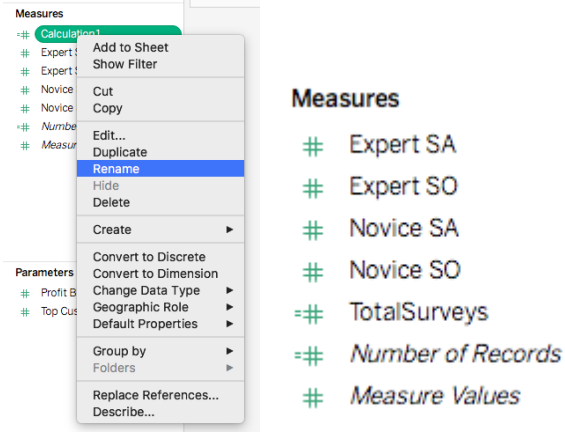


12. **SA** columns indicate that the surveyor reported the ***S**pecies and their **A**bundance*, while **SO** columns indicate that the surveyor ***O**nly reported the **S**pecies* they saw and not the abundance level. **Name** refers to the name of a coral reef that was surveyed. Code is a unique number for each reef site.
    k. Let's figure out whether experts or novices are providing data for each reef. Make a side-by-side bar plot that compares the total (Sum) number of Expert SA reports to the total (Sum) number of Novice SA reports for each reef site.



    l. Let's assume that experts provide more accurate survey data than novices. Are there any reef sites that appear to have significantly more data from novice surveys than expert surveys? If so, analysis regarding these sites may be incomplete or misleading! List any such reef names (hint: ~3-5 locations).
    m. Side note: On average, experts report 80.86 species per survey and novices report 55.64 species – this could bias the overall data.
13. Let's create a calculated field that counts how many total surveys are available per location, regardless of the person. To do this, we will add a new measure that simply adds the ExpertSA, ExpertSO, NoviceSA, and Novice SO columns.
    n. From the top menu, select Analysis >> Create calculated field, and add the equation.



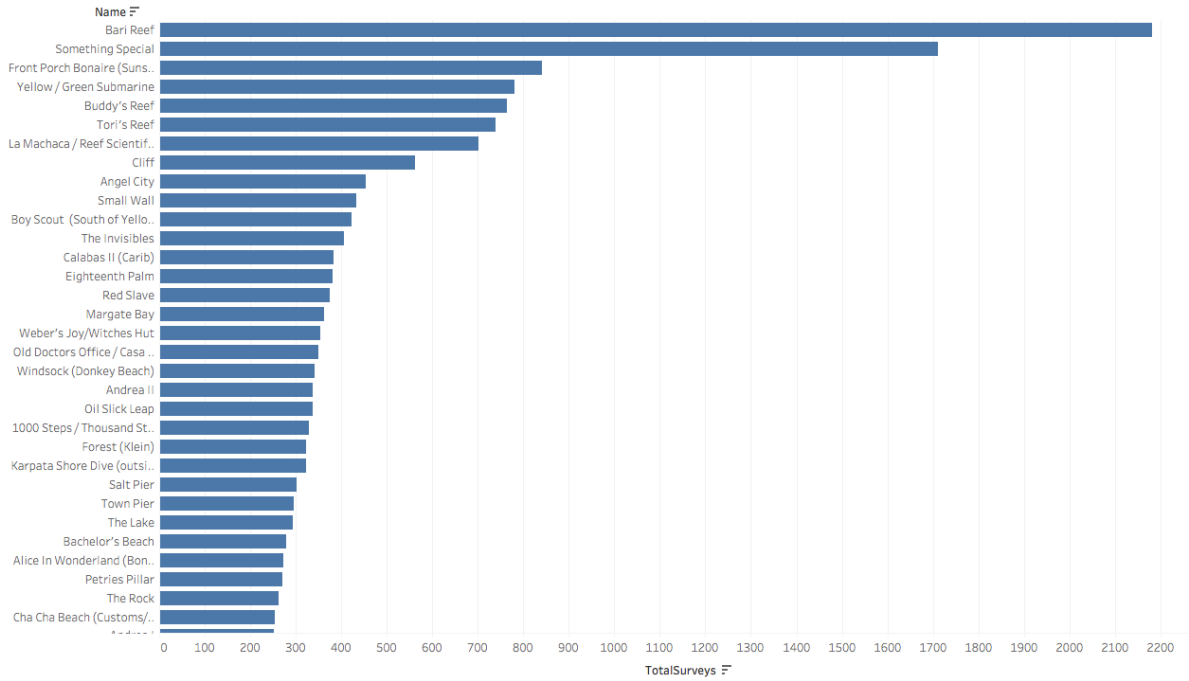


o. Rename Calculation1 to TotalSurveys.

p. Create a sheet of the *total surveys* per *reef name* using a chart type of your choice, e.g., bar.

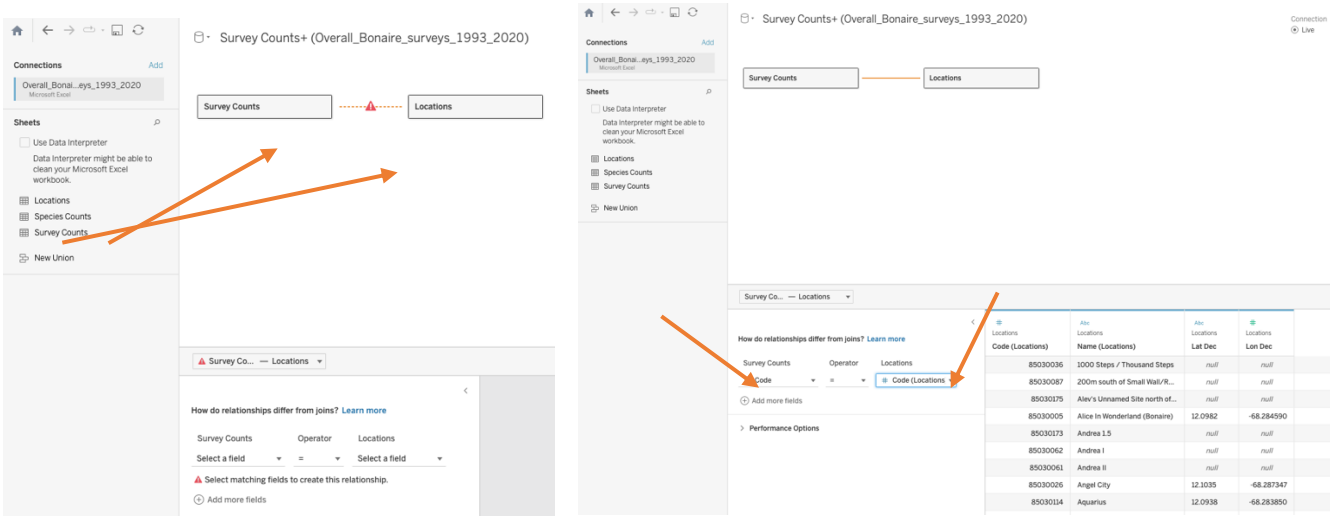
q. Sort the reefs by deceasing numbers of TotalSurveys in the top menu. This menu option can sort ascending, descending, or transpose (flip the rows and columns).

r. What are the names of the top seven reefs by survey count? Were any of these locations identified as being possibly biased due to too many novices from question 10?



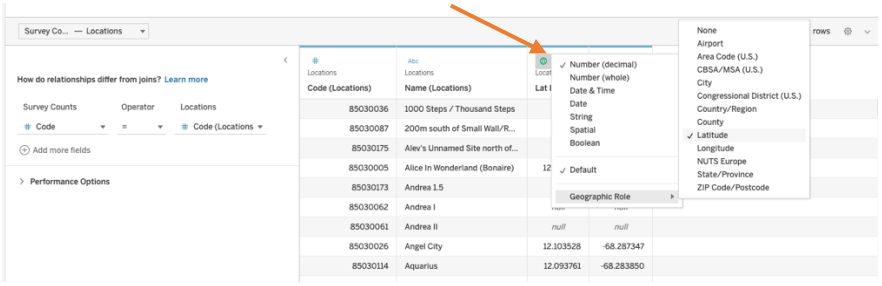
Next, let's place these datapoints on a geospatial map.

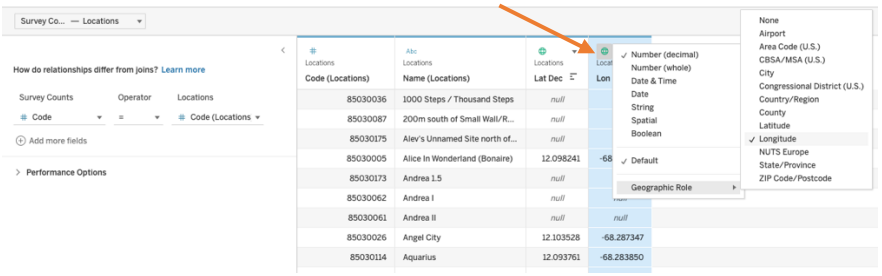
14. Add a new datasource out of the Excel file's '*Locations'* worksheet. Open the Excel file as a new data source. Drag both *Survey Counts* and *Locations* worksheets into the top pane.



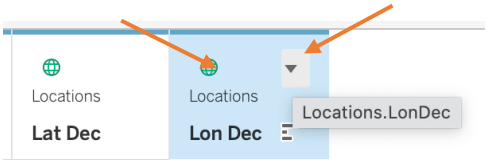
15. Rows from two separate datasources can be merged into a larger row as long as they both contain an identical identifier that can be matched. We are going to join the survey data for each reef location (a row in the *Survey Counts* table) with its actual Lat/Lon coordinates (from the *Locations* table). Edit the relationship to force Tableau to only match rows from the *Survey Counts* and *Location* datasets together if they discuss the same location *Code*. Each reef site gets its own unique *code* and *name*. In the database world, this is called in inner join, natural join, and equijoin.

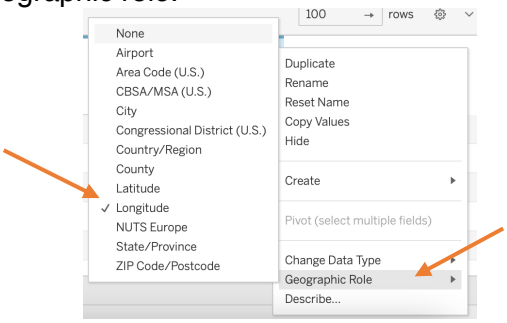
16. Edit both *Lat Dec* and *Lon Dec* columns so that Tableau interprets these as decimal numbers and as latitude and longitude data points. Also, select to use them as columns with actual Lat/Lon coordinates. You also may have to convert them between discrete/continuous/measure/dimension.



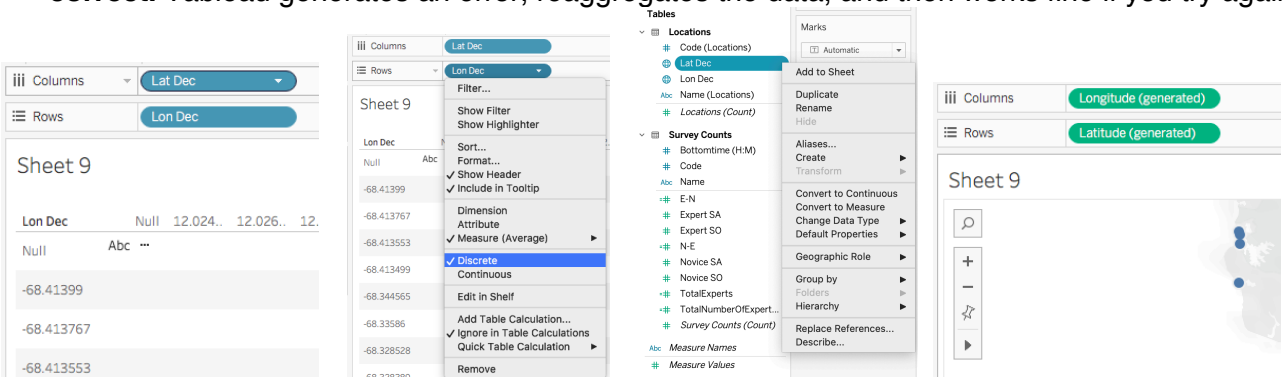


On Tableau Cloud, the menu may be very slightly different. You must make both fields decimal numbers and also assign them a latitude or longitude geographic role.

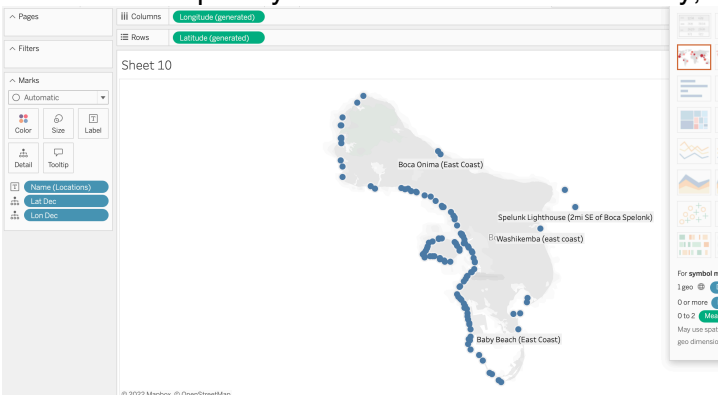




17. Create a new sheet for our map from our new *Survey Counts+* dataset. Drag *Lat Dec* and *Lon Dec* to the chart. **You may have to edit the two pills to convert them both to Discrete if the map is not correct.** Tableau generates an error, reaggregates the data, and then works fine if you try again.



18. If Tableau interprets your lat/lon coordinates correctly, it will generate a map with a point for each row.



19. Add *Expert SA* to set the size of each point and *Name* to each point's label. Hint: On the left Tables pane, you may have to select the *Lat Dec* and *Lon Dec* dropdown menus and *Convert them to Discrete*. When you add *Expert SA* to the *Marks* pane, you may have to change *Measure(Sum)* to *Attribute*.
20. Where do experts seem to conduct the most surveys?
21. Reflect+answer: Why might the Southwest coast of the island be heavily surveyed and not the East coast? Could this bias the type and abundance of species within the surveys of this trusted dataset?