STATISTICS TEACHER
SUPPORTING THE TEACHING AND LEARNING OF STATISTICS
STatistics Education Web STEW
online journal of K-12 statistics lesson plans

Using Technology to Create Data Elements

Leticia Perez, University of California at Los Angeles Center X
Published: October 2021

Overview of Lesson

In this lesson students explore how photographs can be processed to provide insight and data about their favorite outdoor space. Following an initial student survey about the details of their favorite outdoor space, students develop methods to collect additional data from the photographs submitted by their classmates.

This lesson was written to demonstrate how students might engage in the Statistical Problem-Solving Process at Level B as detailed in the *preK-12 Guidelines for Assessment and Instruction in Statistics Education I* (Bargagliotti et.al., 2020). This lesson also connects to two other STEW lessons written by Sheri Johnson and Pip Arnold, both of which explore different ways teachers might help students use data to tell stories about their favorite outdoor spaces. *Using CODAP to Tell Different Stories* introduces students to the overall statistical investigation process and highlights how to use CODAP for data interpretations. *Questions Throughout the Statistical Problem-solving Process* emphasizes the role of questioning throughout the investigation process.

Type of Data

- More than two variables
- Data generated or collected as a class

Learning Objectives

- Students use algorithmic thinking to plan a data collection method and quantify features within photographs
- Students describe the variability within data collected from photographs
- CCSS.MATH.CONTENT.7.RP.A.2 Recognize and represent proportional relationships between quantities

Audience

- Students at Level B from GAISEII (approximately grades 7-9).
- *Prerequisites:* Prior to this lesson, students should have experience with division and be familiar with percentages and creating exploratory plots via Common Data Analysis Platform (CODAP)

Time Required

Four 50–60-minute sessions are required to complete this task. Students need to take photographs of their favorite space and submit to a shared digital folder after the first class session.

Technology and Other Materials

- *Technology:* The process detailed in this lesson would require students to have access to a phone (or a camera and computer) to submit a photograph of their favorite space as well as the ability to complete a Google survey.
- Computer for each pair of students with Internet access for CODAP and Google Slides.

Lesson Plan

The teaching sequence is likely to take three to four 50-60-minute class sessions to complete. Possible breaks in the plan are indicated, but teachers should plan to take a break with their class where it is appropriate as some activities will take longer with some groups and less with others.

SESSION ONE
Setting the Scene: 5 minutes

The teacher opens the lesson by describing that she has been watching documentaries about life on other planets. Watching these documentaries made her think about what it would be like to live on Mars and how she would miss being out in nature and playing her favorite sports. This wondering led to the overarching question for the lesson series, "What features of an outdoor space make it special or feel just right?" She then shares a photograph of her favorite outdoor location and why it is important to her. The think aloud serves as a stimulus to invite students to ask follow-up questions about the space and why it is important to her.

After students' follow up questions, the teacher introduces the lesson's overarching focus question to the class, "How do our favorite outdoor spaces compare?" The overarching question is used to initiate students into some predictive thinking around the types of primary data that might be useful to collect, explore, analyze, and interpret. The open nature of this "setting the scene" is to widen student perception about how data might be collected and what we consider data.

Posing Statistical Investigative Questions: 25 minutes

In this activity, the teacher prompts students to consider the limitations of data collected through a journal entry.

1. Ask the students to close their eyes and picture their favorite outdoor space. Provide students two minutes to journal in as much detail as possible the space they thought about. The teacher asks for two to three volunteers to share some details about their outdoor space. Students then respond to the prompt: *We have collected information about our favorite outdoor space by writing in our journals, what are other methods we could use to collect information about our favorite outdoor spaces?*

2. The teacher then breaks students up into small groups of 3-4. Firstly, students compare details about their favorite spaces, and then are asked to brainstorm two to three other methods they could use to collect information to better understand the favorite spaces of everyone in the class. Prompts to use could include:
    - *Were there any patterns you noticed across your favorite spaces?*
    - *How easy or hard is it to look for patterns or things in common within your journal entries? What if you had to look at an entire class's journals?*
    - *Other than journaling, what are other ways we could share details about our favorite spaces?*

3. The teacher brings the small groups back together and asks students to share their thinking. The class discussion is summarized on the board. The goal is to guide the discussion to help students identify that surveys and photographs/videos would help us better understand what the favorite space looks like and that surveys are a tool to collect information that is similar to a journal but responses are shorter and organized by survey question.

4. The teacher highlights the themes noting that survey questions are useful ways to collect the same information from a group of people. Digital surveys can also allow participants to upload video or photographs. The teacher asks: *"How might a photograph help us explore our question,* **How do our favorite outdoor spaces compare?***, more effectively than a question about* **How often you visit the favorite outdoor space?***"*

The method of data collection for this activity uses a common spreadsheet where the answers to data collection questions are stored. If you were interested in finding out more about conducting a survey with students about favorite outdoor spaces see *Questioning Throughout the Statistical Problem-Solving Process* and *Using CODAP to Tell Different Data Stories.*

5. To introduce the students to language that could be used to compare two photographs the teacher adds a second photograph next to her favorite outdoor space shared during setting the scene. *If we wanted to compare these two photographs, what questions might I ask myself?* The teacher asks the students to use terms from their word bank (e.g. compare, differ, vary, usual, typical, how often) to create some initial questions. Examples of possible student generated questions include:
    - How does the amount of sky compare between the two photographs?
    - How does the number of trees differ between the two photographs?
    - Is the photograph in the forest "greener" than the photograph taken by the lake?
    - Are there more man-made structures in the left-hand photo compared to the right-hand photo?

6. The teacher records these questions in a central location for the class to see. She notes that some of the questions can be answered directly by the photographs and some cannot. Generally speaking, statistical investigative questions ought to meet two criteria: can the data (photograph) be collected and submitted by the students and do we anticipate that there will be variation in our data (photographs)?

Many of the questions posed during this phase are not formal statistical investigative questions. This stage of the lesson is to help students identify variables of interest and support students with the necessary language to pose future statistical investigative questions. See additional notes at the end for a more detailed description about statistical investigative questions.

Collect/Consider Data Part 1: 5 minutes

1. For homework, or before the next session students will find or take a photograph that best represents their favorite outdoor space and meets the following safe photo rules (listed below).
   - Rule 1: Permanence: Photographs you share on the internet will be visible to the entire class. You should presume that anyone can look at the photographs at any time.
   - Rule 2: Identity: Do not share photographs with identifiable information (house address, recognizable faces, personal information that can be used to identify someone, etc.)
   - Rule 3: Safe and appropriate: No selfies (this is redundant with rule 2, but kids will be kids) since we are studying the space not the people within it.

2. Students email the photograph to their teacher by a set time/date so the teacher can prepare the photographs for the next session.

SESSION TWO
Collect/Consider Data Part 2: 45-60 minutes

Background information about using quadrats
Photographs and video can be a source of categorical and quantitative data for students to analyze. This section is devoted to developing methods and data collection questions for the class's photographs. This method can be used for many collections of photographs collected by students and is inspired by the quadrat method used by scientist shown in Figure 1 to quantify and more reliably collect data for specimen diversity within a given area.

Figure 1: Quadrat method (*Using Quadrats - FSC Biology Fieldwork*, n.d.).
Photo Credit: Anne Maben

Prior to the lesson preparation for the teacher

The teacher imports all the photographs into a collage on slides. Then he or she will create a 10 x 10 grid on a slide using the insert, table feature. The teacher imports the student photograph next to the grid. The process is repeated until all students' photos are in the deck.

Figure 2: Slide with photo collage and individual student photo and 10 x 10 grid.

The slide show will have one slide that is the photo collage of all the class photos, and this will be followed by each individual student photo placed alongside the 10 x 10 grid.

Class activity

1. The teacher introduces the slideshow with the student collage photographs and reminds the class of the overarching question, *"How do our favorite outdoor spaces compare"*

2. In small groups, students are given nine minutes to look through the slide deck and record any patterns, observations, wonderings they have about the collection of photographs.

3. Next, students identify an attribute (variable) within the photos that they feel might connect to the overarching question of *"How do our favorite outdoor spaces compare?"*. The students then sort the photos into two or more categories based on their attribute (variable).

    - Note: It is possible that the initial photographs shown by the teacher may affect the types of categories and photographs submitted by students. Sharing favorite locations within the community the school is located provides an excellent opportunity for the teacher to communicate an awareness of the student's daily lives as well as for the teacher to learn interesting new locations that are important to his or her students.
    - Students might note the orientation or size of the photo, subject matter, or activity within the photograph. Encourage and accept all ideas as they could lead to an interesting exploration.

4. Groups share out their observations, patterns, and categories to the whole class.
    - The teacher records the categories (students might not recognize the variable they have used but can give the categories that describe the variable – column 1 in Figure 3) or attributes (variables – column 2 in Figure 3) identified by the class.
    - The teacher then probes the group's thinking to identify the variables that their categories describe, or categories/values for their variables.

- The teacher then asks what they might describe, count, or measure for the variable. (See Figure 3 – column 3).
- When considering what we might count or measure, students may initially create informal conjectures such as, "pictures with water will be blue, so we can count the amount of blue."

Categories (or values for quantitative variables)	Variable (sometimes called an attribute)	What we could describe, measure or count.	A question we might be able to answer…
Student gives: Nature versus human made	Generates: Nature: lots of trees, water and sky Human: we can see human made stuff or activity	Generates: We could count the number of trees in each photograph. We could use a survey to see why students chose this space. We could "measure" the amount of water in the photograph.	Generates: Do students in our class tend to choose natural spaces rather than human made places as their favorite outdoor space? Do our class photographs generally have more natural features than artificial features?
Generates: Green, blue, brown, purple, grey, black, red, yellow, orange, white	**Student gives:** Predominant color of photograph	Generates: We could find out what area of the photograph each color is. We could describe the color we see with our eyes when we look at the photo.	Generates: Do the class photos of sea, lakes or rivers tend to have more blue in them than photos of forests and parks?

Figure 3: Exploring categories and variables in photographs.

5. The teacher poses the question, "*If there is a photograph that is unclear as to the data we record, or data recorded that is unclear, what can we do as statistical investigators?"*

   Students will likely identify that there is an opportunity to ask classmates for clarification or details of the photo. This is an advantage of working with primary data that the students are familiar with. Anticipating and planning for messy data ahead of time can help improve the data collection process.

6. The teacher revisits the third column in Figure 3 and presses the class to consider, "*How might we turn each of our photographs into data that we could input to a data table*?"
   - Responses might include: *We can use words like few, some, many trees to describe the photos (categorical) or we can count trees.*

- Then the teacher asks the students to look over the photos and asks, *"Would counting each tree be possible for all of our photos? Counting object in a picture is one way we can create a numerical variable, but it might be difficult to "count trees" for photos taken in a forest."*

7. The teacher shares a photograph of the quadrat method shown in Fig 1 (*Using Quadrats - FSC Biology Fieldwork*, n.d.). Scientists use this method to quantify and more reliably collect data for specimen diversity within a given area and wonders aloud, *"How might we do something similar with our photographs to capture numerical data for some of our variables of interest?"*

8. Within the slide presentation, the teacher shows figure 2 with the grid and shares, "rather than counting individual trees we can count the number of boxes with trees." The teacher models the following procedure for several photographs described and depicted below.
    - Rule 1: As much as possible, keep the aspect ratio (ratio of width to height) of the original photograph the same. We can make the photo larger or smaller by dragging from the corners but not the sides, this keeps the height and width in the same proportion.
    - Rule 2: We can make the grid larger or smaller, as long as we keep 100 squares. For many landscape photos this process is the fastest (see Figure 4)
        - Step one: Import one photo per slide.
        - Step two: Drag the photo from the corner increasing the size so that the photo has filled most of the slide.
        - Step three: Select insert, then table 10 x 10 to create a grid.
        - Step four: Resize or drag the grid so that it covers the entire photo

Figure 4: Photographs depicting the steps describe to resize the photograph and grid to maintain overall dimensions.

9. Next, the teacher asks the students, *"How will we use the squares to track the number of trees or other attributes of interest?"* (Figure 4). The purpose of this stage is to develop a shared class data collection approach, as data collection methods need to be applied consistently. There are many data collection questions a group working together might need to explore before collecting the actual data.

Questions to explore with the class about the data collection method might include:

- *Is it possible for multiple attributes to be counted within a single square?* In the case of the sample photograph above (Figure 4), the 3rd row from the top is filled with both trees and sky. Students will recognize that multiple attributes might be present within a single square and as a class there is a need for a uniform approach to collecting data. E.g., do we count ½ or 1/3 squares?
- *What advantages might a 10 x 10 grid offer?* We can resize the grid, maintain the original photo dimensions, and report the variables of interest as a percentage. The percentage reporting allows us to ignore the size of each photograph.
- *How might our data change if we do not maintain aspect ratio to fit into this grid?* If we change the dimensions of the picture and do not keep it proportional certain features might be under or over-represented.
- *Now that we know our data method collection, which attributes or variables from our class chart will we record? What would a statistical investigative question be?* Gather ideas for statistical investigative questions, ensuring that the variables, intent, and "group" are clear. E.g., Do the class photos of lakes or rivers tend to have more blue in them than photos of forests and parks? – this is a comparison investigative question (intent), the variable is the amount of blue in the picture, the groups are: class photos of sea, lakes or rivers and class photos of forests and parks.
- *If we want to compare nature vs human made, what might make sense to count?* At this stage, the teacher guides the students to identify the attributes or variables of interest. E.g., if we are wanting to compare the amount of blue in a photograph then we need to count how many squares contain blue.

10. Each pair of students should generate a statistical investigative question that they want to answer using the data cards (slides). If you want everyone to use the counting squares method, then ensure that to answer their statistical investigative questions this is the data collection method needed.

11. Each pair of students should identify the counts that they need from the photographs and collate all the different counts onto the board or a similar shared space. E.g., across the class they might need to count how many squares are of each color (Green, blue, brown, purple, grey, black, red, yellow, orange, white); other groups might want to know about how many squares contain trees, water, sky, man-made objects; yet another group might be interested in the dominant color of the photograph.

12. Once all the required information is identified and listed, direct students to the class spreadsheet where they can input their information for their individual photograph. This could be a Google Sheet. The columns would contain all the different variables, the rows contain the information for each photograph, this is probably identified by the student's name or initials.

13. Students work in pairs to collect the square counts for the identified attributes of their personal photographs using the grid method. The pair will also agree on the dominant color within the photograph (e.g., black, grey, brown, blue, purple, green, red, yellow,

orange, white). The counts and color names for each of the students' photos are added to a single class spreadsheet.

Trees	water	man made	sky	dominant color
14	25	0		blue
28	27	0	28	green
43	9	0	15	green
26	0	0	40	brown
47	0	0	8	black

Figure 5: Spreadsheet of data input for each photo.

14. At the end, the students complete an individual journal reflection, *"What successes and challenges did you and your partner experience collecting data today?"*

SESSION THREE
Collect/Consider Data Part 3: 25 minutes

Reflection from previous session
While working in pairs, it may have been difficult for students to agree on whether to count a square or what the dominant color for the photographs is. Negotiation of data collection methods is meant to provide students with insight around the many decisions and opportunities for bias or error that are created during the data collection period. Some of this error may be intentional E.g., team decided not to use a particular photograph, and others unintentional, E.g., miscounting the number of squares for a variable. Furthermore, students often hold informal and unspoken hypotheses that shape the data collection or interpretation process. For example, some students might predict or assume that "natural" photos are more likely to be blue or green vs grey or black. Such predictions or wonderings could lead to new investigative questions, data collection and transformation procedures.

The next section details a process to explore the statistical investigative question, "*How might the way we characterize the dominant color within a photograph compare to the way computers "see" the dominant color?"* To answer this question we need to collect more data.

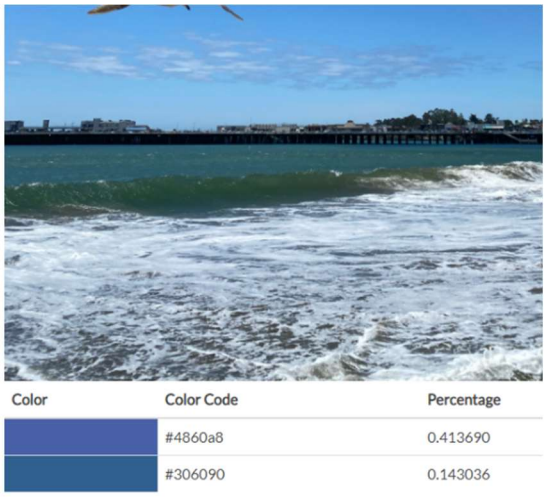
Using digital tools, we can collect additional data from photographs that once were unavailable.

Data collection method
The following steps describe how classes can use a web-based application to analyze the predominant color within the pixels of a photograph (See Figure 6). Each photograph is composed of individual pixels each assigned a singular color. This web-based application and others like it, tabulate the frequency of different pixel colors (reported here as a hex code) within the photograph. In this lesson, we have named the first two values returned as "primary color" and "secondary color". Hex code values will display as color within web-based applications such as CODAP (See Figure 7).

1. The teacher first demonstrates how to use a photo color analyzer and record the data:
    a. Visit the website: http://www.coolphptools.com/color_extract

b. Scroll down below the demo photograph to upload the photograph and in the "Number of colors" field enter two. Leave all other settings as the default setting.
c. Click Run. The site will return something like Figure 6.
d. Record the primary and secondary "color code" and the percentage of the photograph for those two colors in the class spreadsheet used in the previous day's lesson.
e. Using the same color name labels from earlier (i.e., black, grey, brown, blue, purple, green, red, yellow, orange, and white) assign categorical descriptions of your first impression of the primary and secondary color in the new columns of the dataset.

Figure 6: Pixel analysis via coolphptools.



Color	Color Code	Percentage
	#4860a8	0.413690
	#306090	0.143036

2. The student pairs upload their photos and collect the primary and secondary color hex codes (referred to Color Code in Figure 6) as well as pixel percentage and then adds the new data to the class spreadsheet used earlier in the sequence. Columns 2,3,5,6 are values from the coolphotos page. The hex color codes will show up as color swatches as shown in column 2 and 5. Columns 1,4,7 are categorical labels assigned to the color by the student pair examining their photographs. This process increases the total number of attributes to the photos to 12. To see a summary table that describes the variable and data collection method refer to Figure 16.

Primary ...or name	Primary ...xel color	Primary ... color %	Second- ...r name	Second- ...xel color	Secondary ...pixelcolor %	Primary ... human
black		0.16	grey		0.15	black
brown		0.2	blue		0.19	black
brown		0.34	black		0.23	brown
black		0.55	black		0.12	blue
brown		0.37	black		0.35	purple
black		0.13	grey		0.1	bronw
brown		0.18	grey		0.17	green
blue		0.2	black		0.17	black
blue		0.41	blue#3...		14	black

Figure 7: Pixel color data collected uploaded into CODAP.

3. The teacher then uses the import feature to upload the shared class data from the spreadsheet to CODAP and shares the file for the students to analyze relationships between the 12 attributes.

During this phase of the lesson students used technology to reveal additional numeric data that was "hidden" within their submitted photographs. Often times, statisticians and scientist may use collected data and mutate or transform it into new attributes such as calculating the speed for objects rolled down a ramp from the distance and time traveled data. It is important for students to also realize that new tools can allow for scientist and statisticians to revisit existing datasets, in this case the survey and photographs, and uncover or more formally describe attributes that were previously unseen or unquantified.

SESSION FOUR

Advance preparation:

Common Online Data Assessment Platform (CODAP) is an open-source data analysis application that helps students generate representations quickly and easily. This web-based software recognizes whether an attribute is categorical or numerical and then selects the best type of representation. Data can be directly input into a table or uploaded from a spreadsheet. Additionally, a teacher can preload a CODAP file with a dataset and plots and then share with the students through the sharing link. Sharing files in this way generates unique copies for students to explore and interpret their data while maintain the integrity of the original file. Students do not need large amounts of experience with CODAP to engage in the data exploration. Additionally, CODAP will display HexData as color bars, see Primary Color in Figure 7 and 8.

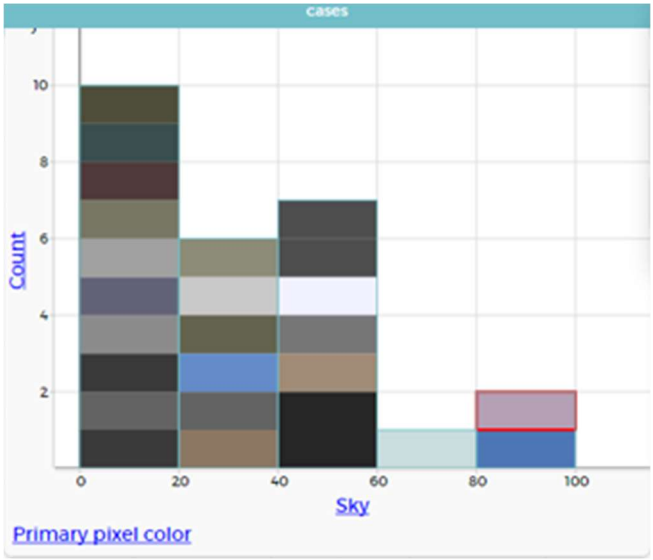


Figure 8: Sample plot generated by students. The x-axis represents the percentage of the photograph that is sky. Each box displays the dominant pixel color for the photograph as indicated by the color and the attribute shown in the bottom left hand corner.

Students will need a shared CODAP link with the class data preloaded into the page. The simplest approach is to generate a shareable link that makes a "forced copy" for any user. This link can be shared via a hyperdoc or email. To create a shared link select: Menu → Share → Get Link to Shared View → Enable Sharing → Copy and share the link electronically.

Analyze and Interpretation: 45-60 minutes for each statistical investigation

Returning to our overarching focus question, "How do our favorite outdoor spaces compare?" The following portion of this lesson plan describes three different statistical investigations that could result from data collection described above. The statistical investigations are presented in some detail to aid teachers in thinking about possible ways they could extend the investigation.

Statistical Investigation 1: Comparing an individual observation case to the group for a variable of interest.

To help students become familiar with their data and CODAP we first being by considering, *"How does my picture compare to that of the class?"* There are several attributes that students could use to address this question i.e., amount of sky, trees, color of photo or human made features. Note that the attributes available for student investigation are dependent upon the variables identified within Session 3 and their student interest.

Formulate Statistical Investigative Questions:

1. The teacher first displays on the board the statistical investigative question that will frame the data investigation. These statistical investigative questions are summary questions seeking to describe general patterns between the student's data and the class.

    a. *How much of our class photographs contain the sky?* This question's ambiguousness is framed in student language and is designed to create a larger variety of interpretations by the student pairs. Lastly, it provides a reflection opportunity for student to revisit their statistical investigative question and refine it. A general question will likely result in different analysis approach. Some students might report the proportion of photographs that had sky vs did not, while others might use the percentage of squares attribute and explore the variability. This variation in student interpretation and approaches provides a rich opportunity to further explore the qualities of good statistical questions. Should you wish to make this exploration more direct you might consider using the statistical investigation questions, "What percentage of each photograph shows the sky, and how do these percentages vary?"

    b. *What is the color for the sky in photographs?*

Collect/Consider Data:

1. Students are given a link to CODAP with their class data preloaded.
2. The teacher demonstrates how to scroll through the table to identify individual cases or observations. In CODAP, an individual observation with all of its attributes can be displayed in one place, this is called a case card. Note: CODAP uses attributes instead of variables. The teacher demonstrates how to use the mouse to drag the attribute onto the graph to create different plots. Additionally, the teacher notes the features which allow students to customize plots e.g. bars, custom colors, functions. To learn more about how to use CODAP, please refer to the tutorials and videos within the resource section
3. Students compare their picture's data to that of the class by writing 2-3 sentences. Possible analysis question prompt: *How your data is similar and different to that of the class?*

Analysis and Interpretation:

1. Next, pairs of students are instructed to specifically focus on the values reported for the attribute sky.
2. Students are given 5-10 minutes to make as many visualizations as they can make with the attribute of sky. An example of the variety of plots that can be generated are shown in Figure 9.

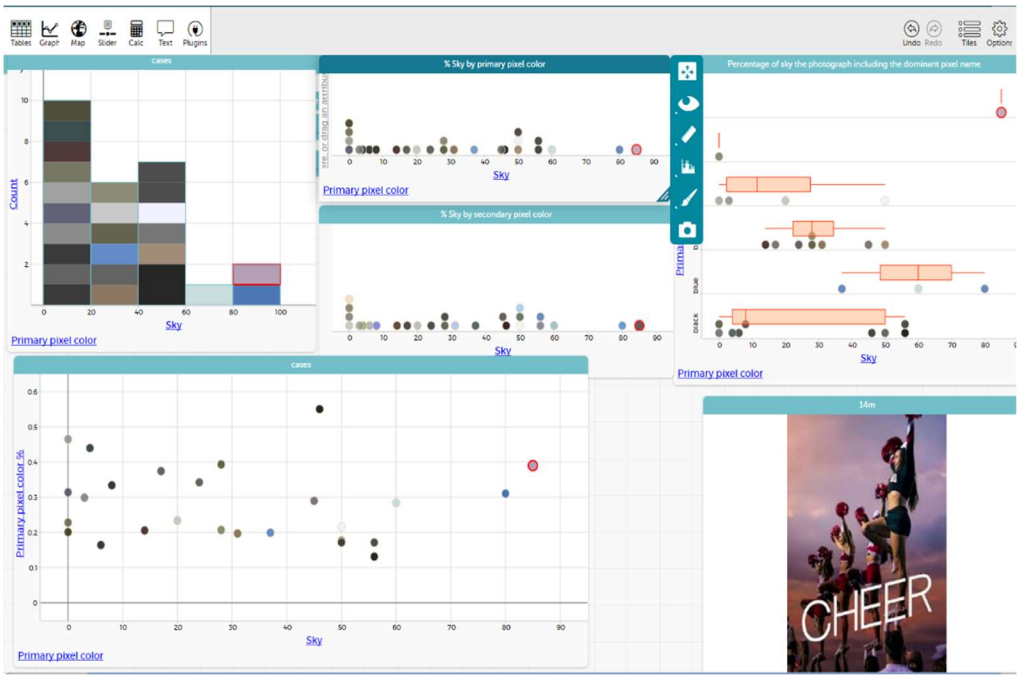


Figure 9: Variety of plots and student photo in CODAP.
A full page view can be found in the appendix or you can view a link in CODAP.

3. Have students scroll through the table to find their photo observation again.
4. Instruct students to click on their observation or case card. This will highlight all of their particular data points within their displays.
5. Next, ask students to record all noticings, observations and or wonderings for two or more representations.
6. To encourage students to develop more elaborate observations and descriptions of their displays, a teacher might offer sentence frames. Sentence frames or starters help scaffold

the academic language often needed to analyze or interpret plots and or help student consider different components of the plot that may be available to describe. Sample sentence frames for data generated within this task are offered below.

Sentence Frames: (provided by teacher)

- My observation is to the right, left or with the most common bin.
- There are more or less observations to the left or right of my bin.
- I do or do not think my observation represents the usual / predicted / typical count because
- Pattern: As _____ happens I notice that _____.
- Comparison: My photograph is similar /different to the rest of the class because
- Comparison: As the percentage of the sky increases the _____ variable…

Sample Student Response for Figure 8:

I noticed that my photo was the only one with lots of purple in it. I wonder why so many of the photos are dark in color? There are more bins to the left of my observation bin. I think my observation of sky color is unusual because most of the bins are to the left and there are only two observations in my bin. As the percentage of sky increases in the photograph black and grey is less common. Two out of 26 photographs were 80% or more of sky; one was purple and one was blue.

Extending the Interpretation Phase:

As a whole group the teacher can support students in developing statistical interpretations of multiple representations for the same variable. Particularly, the use of visualization software applications highlights how the same data can be represented in different ways and this variety foregrounds certain patterns and trends. Some of these discussion prompts might include:

- *Would we write the same captions for each of the plots? Why or why not?*
- *How might some plots make it easier to tell the data story of this student's picture than others?*
- *How might our interpretations of data be influenced by our personal data that was contributed to the class set?*
- *If my picture was mostly grey colored concrete, how might the questions I wonder about or the captions I write vary?*

Offering a diverse set of visualizations for a single variable also provides an opportunity to frame plots as a data storytelling tool. All these plots can be used to help students describe how their picture compares to the class.

- The boxplot (Figure 10) allows students to formally describe how the primary pixel color varies in terms of assigned name, "Primary pixel color name" and the actual primary color appearance displayed within each point. Figure 10 also clearly shows how the percentage of sky in the picture varies within the primary pixel color.

- The case plot (Figure 11) better communicates the variation between primary pixel colors on a visual basis. Examining both plots side by side allows students to test their initial analysis and interpretation intuitions.
- A student might notice that percentage of the photograph that is sky, for (primary pixel color) brown photographs are generally clustered closer together compared to other colors (see Figure 10 and 12). When we look at Figure 11, many of the bars appear to be similar in length for (the primary pixel color) brown which results in a compact boxplot.
- The cluster dot plot (Figure 12) allows students to estimate where the midpoint for the percentage of the photograph that is sky would be without needing to look at both the points and the barplots. The teacher might pose the question, "where on the x-axis might half of the brown points fall on either side." This plot allows us to eyeball the midpoint and can highlight the value. Figure 12 might serve as a scaffold to help students to formally describe brown using the boxplot by stating, "The median percentage of sky for photographs with the most frequent pixel being brown was 30%.".

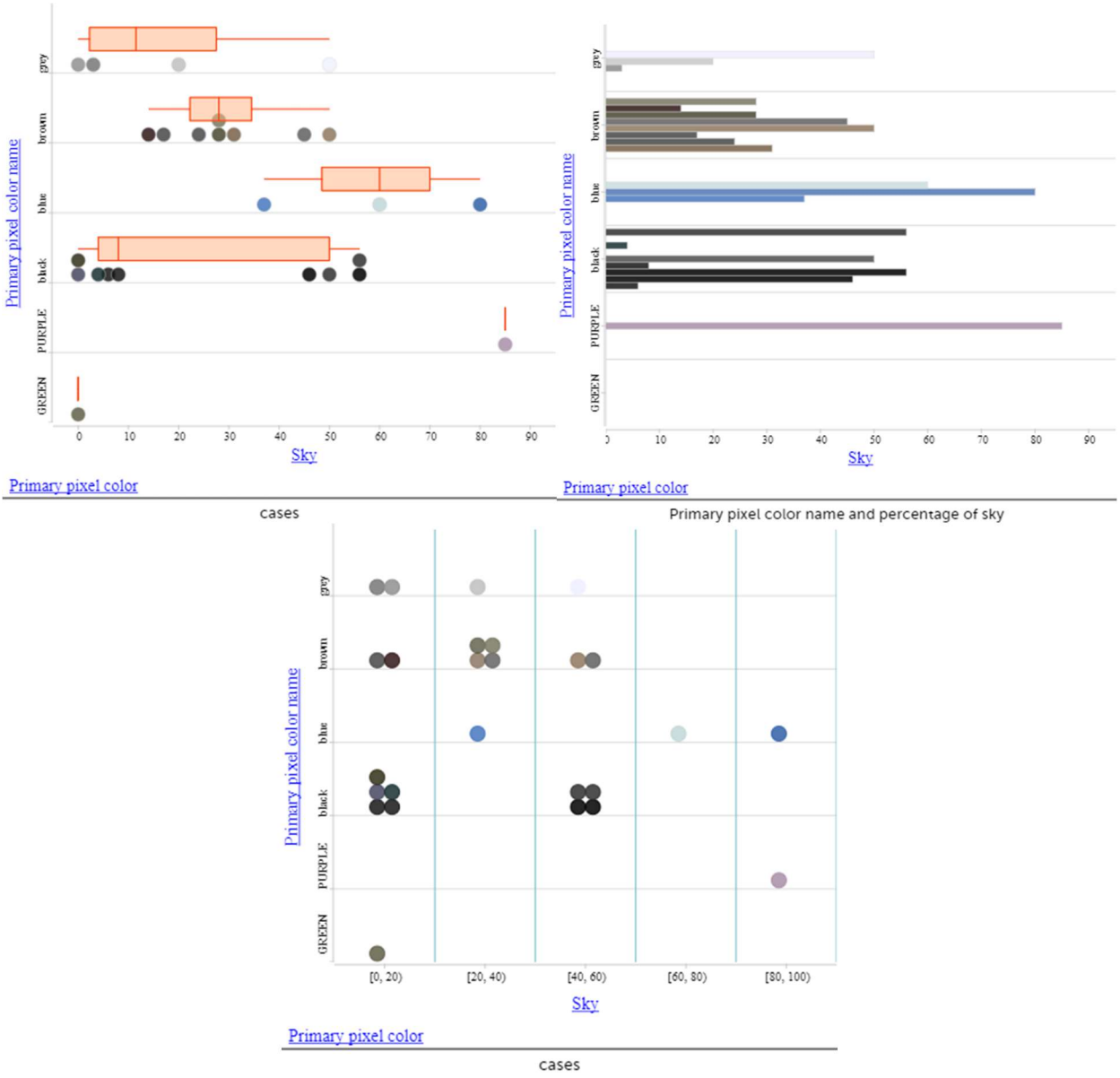


Figure 10, 11,12 : Different representations generated from the same variables of interest via CODAP.

The analysis and interpretation can be further enhanced when we compare pixel color black to brown.

- The very short and very long bars in Figure 10 and cluster of points in Figure 12 provide the opportunity to develop stronger notions of range within a set of values. In this case, students might state that the range of percentage of the photograph that is sky for (primary pixel color) black was 55% while the range of percentage of the photograph that is sky for (primary pixel color) brown was 38%.
- Again, students can visually estimate the location of the median using any of the figures, but some are more intuitive while others are more precise. The median percentage of sky within photographs with black as a primary pixel was 8%.

- Interestingly, the four points of black in the third bin of Figure 12 (40-60%) should inspire additional wondering questions that may inspire further data collecting or statistical investigative questions, *"Was the sky black or was there something large that was black? What is happening in these photographs? Where were the photos taken and were the pictures taken at night? Did the photographs have lots of light and make a strong black silhouette?*"

Providing multiple representations, separately and then side-by-side provides much needed opportunity for students to pose new data collection and analysis questions themselves. This extension highlights the iterative nature of the statistical problem-solving cycle. There are concrete right or wrong answers to questions such as "What is the median percentage of sky for photographs with blue as the primary color", in this case the answer is 59%. This lesson hopes to encourage teachers to use overarching focus questions that are broad and connected to the class' interest or lived experiences. In this way, we invite students to go beyond looking for answers to prescribed questions and to continually interrogate plots, questions, their thinking, the data, and the stories they tell about it. By centering tasks around data storytelling with messy data, we shift the activities' focus away from a binary "right or wrong" task towards a critical thinking task that invites students to use statistical interpretations as tools to describe the world around them.

Statistical Investigation 2: Two or more variables, introducing association.
In this example the students are given a statistical investigative question that was inspired from their initial wonderings. The statistical investigative question detailed in this section aligns to the instructional goals of the teacher which is to informally use "best fit" lines to describe patterns and the relationship between two numerical variables. This investigation centers around students' wonderings about whether photographs with nature, i.e., trees or water, were more or less likely to have human made features.

Formulate statistical investigative questions: *Is there a relationship between the percentage of trees in our class photograph and the percentage of human made features?*

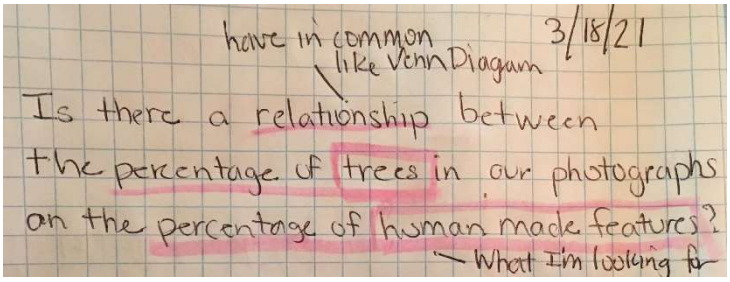


Figure 12: Marking up our Statistical Investigative Questions.

The practice of "marking up" of statistical investigation questions helps students identify and clarify the variables of interest.

- What might "the percentage of trees" in our photographs mean to different readers? Are we analyzing only one photograph and the amount of trees in that photo? Are we looking at the percentage of the total photograph that shows a tree?
- How might we make this question more specific to the data that has already been collected?

Collect/Consider Data: Students underline or box terms or phrases they see as important within the statistical investigation question (see Figure 12). The mark-up helps students identify the variables of interest within the statistical investigative question and to consider what types of plots will best serve the purpose of responding to the investigative question and communicating their thinking. Students are guided to select at minimum the two variables within the statistical investigative question: Trees and human-made features.

Analyze the data:
In pairs, students are given 7-10 minutes to drag and drop the variables along different axes to generate their plots. Students are instructed to look at the variables independently (see Figure 13) and to describe their noticings first.

- Example: *"I see that there are 10-11 photographs that do not have human made features in them or there are a group of photos that have less than 10 squares with trees.*"

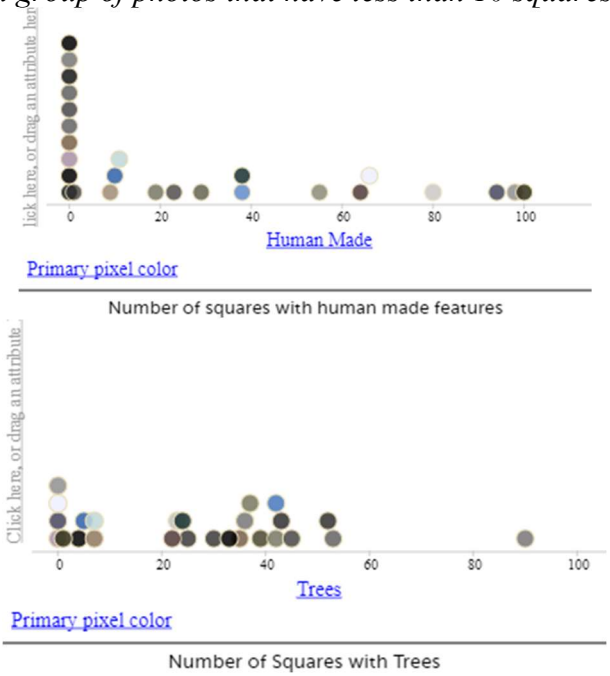


Figure 13: Single variable plots with primary color subset.

Students then create their visualizations with both variables represented in the same plot. Students use noticing sentence frames to mark up the graph in an application such as google draw (See Figure 14). Relevant noticing frames might include:

- *"As the number of _____increases I notice that the amount or number of _____ (stays the same/increase/decreases/no overall pattern).*
- *As I walk up the y-axis I notice…*
- *There are a group of points at ___ and I wonder….*
- *There is a point all alone and it makes me think.... "*

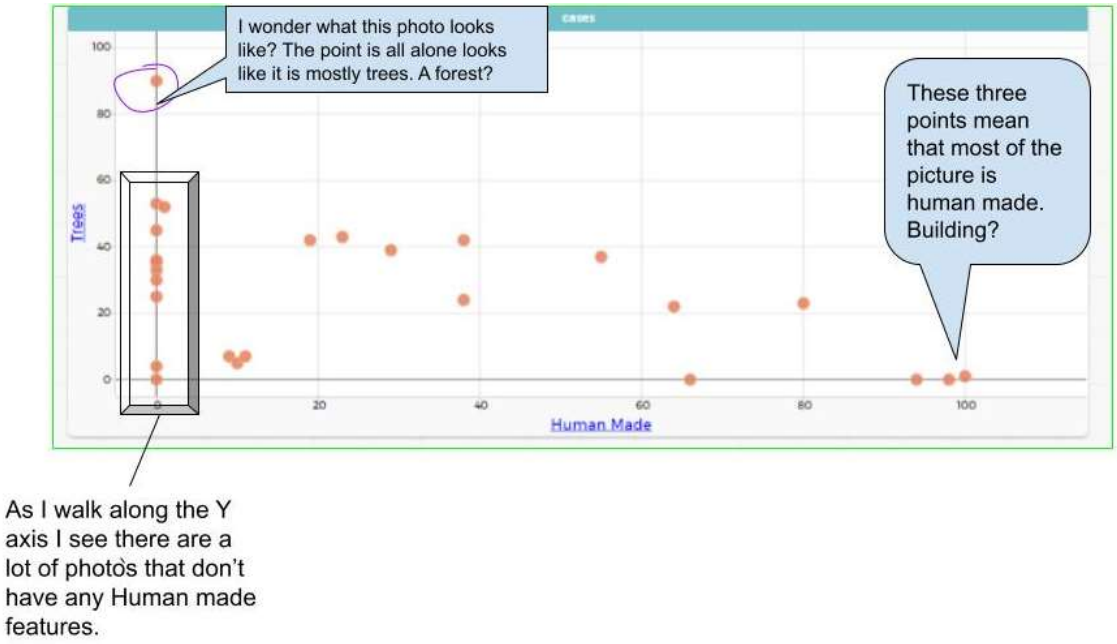


Figure 14: Marking up our Plots.

In pairs, students attempt to write a caption for their plot that expresses the relationship between the two variables. Students then include additional sentences as to how this relationship is a response to the statistical investigative question. Some students may estimate the position of the line while others may choose to use the Least Squares Line built within CODAP (Figure 15).

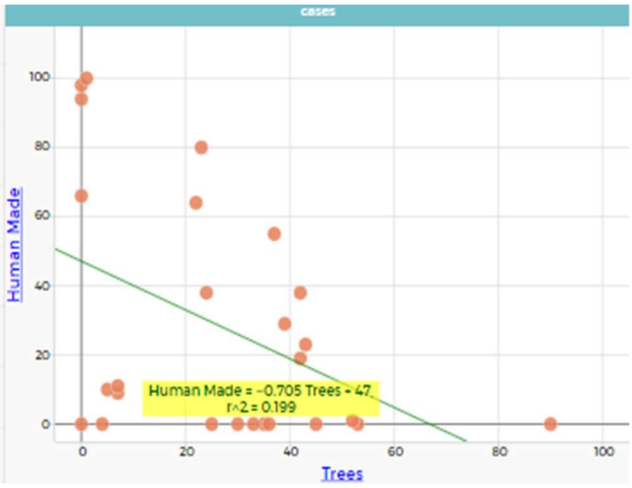


Figure 15: "*Best Fit" line for our investigation.*

As there are many formal and informal ways to describe the relationship between variables, the framing of this task moves away from "answer the question" towards "write a caption in response to the questions." This subtle shift invites students to draw on a wide range of academic vocabulary and explore the value of different techniques.

Next, students explore variation within captions and plots by sharing their products. In this case, a shared Google Doc or digital whiteboard space such as Padlet would easily allow for students to post their plots and captions in a central location for the teacher and classmates to share. An alternative to marking up the graphs digitally is to have students use a template such as a Patterns, Observations and Wonderings Chart. With this template, students copy and paste their display in the top space and then mark up both the plot and capture their thinking. This template can then be shared with other pairs for examination and feedback. Both approaches provide students an opportunity to view the variety of representations from their classmates. Perhaps they might discover when comparing plots that trees on the x or y axis both produce a negative slope line.

The variation within this task provides an opportunity for the class to consider there are many ways to construct an explanation to a [simple] statistical investigative question and even more ways to address their overarching question, "How do our favorite outdoor spaces compare?".

> Student Sample: *There is a pattern in our plot. As the number of squares with human made features increases the amount of trees goes down. In our plot above you can see that half of our points are above or below the line. There are many photographs without any human features at all and we wondered what those photographs were of. We noticed that many of the pictures had forests in the background that took up most of the photo. We noticed that human made was a big category compared to trees. Roads, buildings, benches, ski lifts all counted as human made. We predict that if we compared human made features to water this line would be similar since it is hard to have water and building together."*

Statistical Investigation 3: Summarizing Categorical Data and Digging Deeper Into Data
Comparing the primary color of a photograph.
Students may have noticed that their notions of color vary from what the computer "sees" and what they might see. For example, we might expect photographs with many plants or trees to be green and photos with water or sky to look blue. In this example, we compare the dominant pixel color determined by the coolphoto analyzer and the student assigned pixel color name for photographs that have a high tree count. This approach could also be used for several other attributes.

This exploration introduces sub-setting to dig deeper into datasets and encourages students to ask more investigative questions, data collection questions, and analysis questions. This activity emphasizes the types of questions we might ask about the data collection process rather than a single posed formal question. Whether students' questions are about the validity of the data collection process, developing future data collection opportunities, posing or refining new statistical investigative questions, students need an opportunity to interrogate the data, their biases and preferences, and their assumptions. This work is made possible when we shift the student work product away from a "question-answer" format.

Statistical investigative questions: *What is the usual color of photographs that have a high percentage of trees? How much variation is there within the labels we used for our color descriptions?*

Consider/Collect Data: Students create a table (Figure 16) to identify the variables within the datasets that may be useful in answering their investigative questions: primary pixel color, secondary pixel color, primary pixel color name, trees). The students next summarize the type of data they have i.e., numerical or text/words/categorical and how these values were generated. The chart is designed to help students develop interrogative questions they might ask about the data itself, i.e., the methods of data collection and how this might impact the representations that are generated. Students should update this chart during the data analysis phase in a different color or font.

Variable	What does this data look like in the table? What kinds of graphs can I use with this type of data?	How was this data collected? What was the method?
Trees	This is a number between 0 and 100. Bar plots, dotplots	The number of squares in a 100 grid that looked like there was a tree in it.
Primary Pixel Color	This is a number that computers recognize as color. It shows as a color on the data sheet. I can put this on top of another graph in CODAP to uncover other patterns.	The most frequent color that was found by the computer counting the number of pixels of each color
Primary Pixel Color Name	Word I can make graphs with these categories along the axis that show squares or dots kinda like a bar.	The color name we used to describe the most frequent pixel color identified by the computer
Secondary Pixel Color	This is a number that computers recognize as color. It shows as a color on the data sheet.	The second most frequent color that was found by the computer counting the number of pixels of each color
Primary pic color	Word	This was what we said was the main color of the entire photo before using the computer to analyze the photo

Figure 16: Connecting variables to our data collection and data table.

Analysis: What is the usual color of photographs with a high percentage of trees?
In this stage, students create and interpret plots that explore the primary and secondary pixel color (Figure 17).

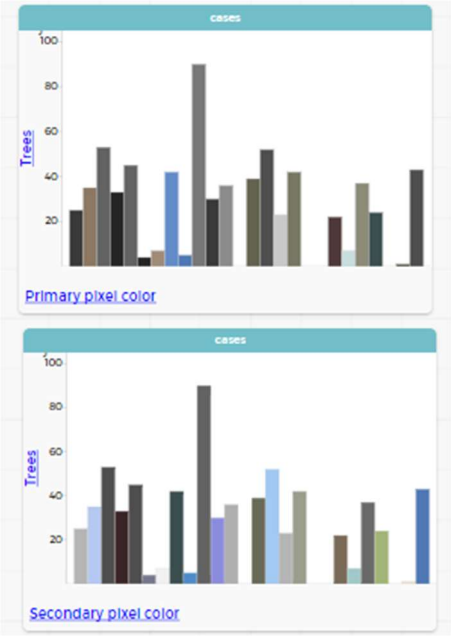


Figure 17: Dominant pixel color.

1. Students are instructed to record the Patterns, Observations, and Wondering that they have when looking at their visualizations. *Sample responses that may result from examining Figure 17.*
    - *Patterns: According to the computer, photos do not tend to show up very green. The most common pixel color for our data appears to be grey or brown.*
    - Observations: *I see that there are a lot of colors that I do not usually think about when I think of trees. There is one picture that has a lot of trees in it, like around 90%. Green only shows up once as a secondary color. Why don't I see green that much?*
    - Wonder: *I wonder if I looked for the third or fourth most common pixels if green would show up more? What color counts as green to the people who labelled it? I wonder what color my classmates would say is the main color in these photos? What colors did we assign to each of these pixels when we uploaded the photos? What if we looked at the colors printed instead of on the screen? From my point of view, that looks like the main color for each picture.*

2. The teacher next poses the question to the class, *"What other attribute might we add to our plots to help us further understand the variation of color in our photographs?"* The teacher tracks the class ideas. These ideas serve to help students further explore the data and consider alternative methods that might help tell the story of trees within our photographs. Possible explorations might include:
    - Exploring the variations within color names that the class assigned to the primary and secondary pixels.
    - Looking at outliers and their case cards to develop a hypothesis as to why the data point is outside the reported ranges.

- Developing another data collection method that might be more accurate. When looking at the variation of the color hex codes displayed as "black" we realize that perhaps many of our color names are inaccurate. The teacher might ask the students to consider whether they should create another attribute and reclassify the data or create additional color name categories. This color variation can be brought to student attention by asking (Figure 18) *How might we preserve our original data set and account for this reclassification? How do we document our data analysis approach to share with other people? Why do we think this variation between color names happened in the first place?*

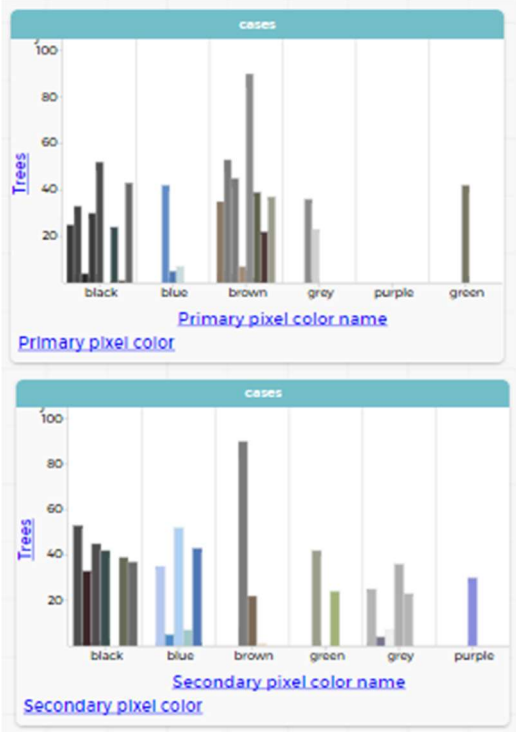


Figure 18: Dominant pixel color by name.

We also might notice that the single photograph that is over 80% trees is classified as brown for the primary and secondary pixel. Students might make predictions as to why green might feel underreported for photographs of trees? What might happen if we increase the number of reported pixels so that can see the top five or ten pixel colors reported by coolphoto analyzer? (Figure 19). *Is it possible that green appears in many different shades and is underrepresented because of our tool or approach?* Perhaps green is commonly found as a tertiary color in photographs.

Color	Color Code	Percentage
	#606060	0.289702
	#484848	0.206488
	#787860	0.195893
	#909078	0.094286
	#303030	0.050774
	#c09090	0.039286
	#a8a8a8	0.025714

Figure 19: Dominant pixel color.

The goal of this activity is to help students ask questions of not only the dataset, but their methods for collecting and analyzing data, to interrogate their process. As a class, it is possible to generate many investigative questions and data collection methods questions. This provides an opportunity for students to expand the questions they ask themselves while engaging in particular statistical investigations and supports the development of critical consumers of data products.

Interpretation: Writing a Data Story

In this next phase, student pairs are tasked with writing their Data Story. When we ask students to write data stories we provide an opportunity to expand the misconceptions about what statistics can "look and feel like". The field of statistics is a tool that can be used to describe the world around them and compelling stories centered around data are a crucial component of statistical and digital literacy. Students are tasked with not only providing an "answer" to their statistical investigative questions but also to connect back to the context that initiated the question in the first place. In this case, our overarching question looked to use data to tell the story about the class' favorite outdoor spaces. Framing the task as "Data storytelling" empowers students to go beyond the data at hand, express the process they used to arrive at their answers, and explore potential implications of their work.

Sample Student Pair Data Story:

We started wondering about our class photographs and trees, Do lots of our pictures have trees? Are there lots of trees in our favorite places? Did we think the places with trees were pretty or fun to visit?

We discovered that the median percentage of photograph that had trees is around 28% (Figure 20).

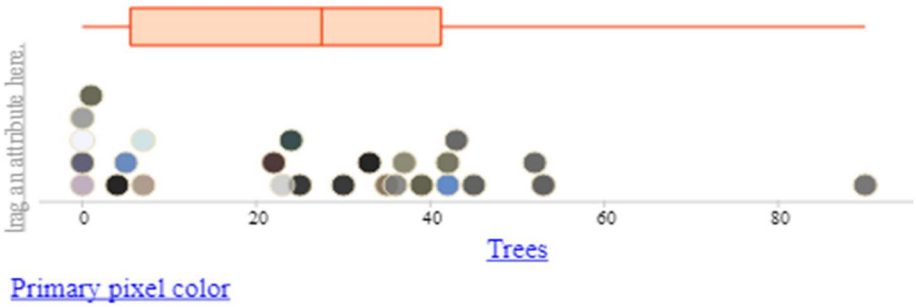


Figure 20: Number of squares with trees by pixel color (percentage).

The lowest number (and percentage of the photograph) of trees (the minimum) was 0% and the greatest percentage of the photograph that was trees was 90% (the maximum). There is not a clear relationship between the color the computer sees and the percentage of the photograph that is trees. The computer did not see the color green very often as the color green is not obviously in the plot (Figure 21). When there were less trees it seemed that the main color was a little bit lighter in color. We were really surprised that we did not see green very often.

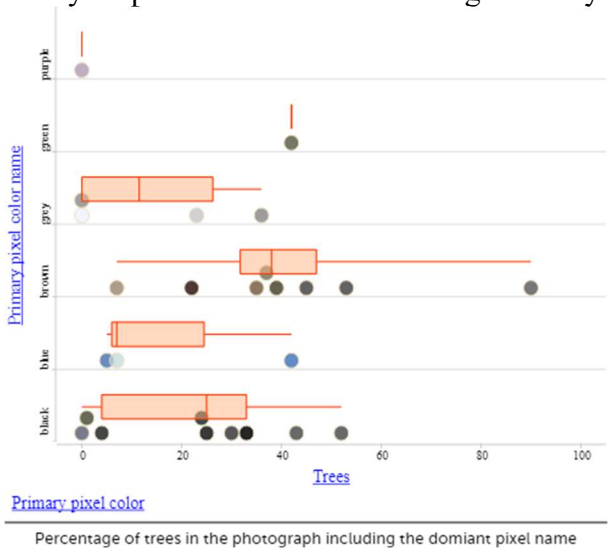


Fig 21: Box plot showing the percentage of photograph that is trees split by color.

The main color that we as a class saw in the photograph was not usually the main color the computer saw. There was a lot of variation in what we called grey or brown. We were also not very good at describing the main color that the computer saw. In this graph you can see that when the computer showed us the most common pixel color and the name we used to describe that color did not always match. We used black and brown to describe the color more often than the other colors. When it came to trees in photographs the computer and our class did not choose green as the main color.

Concluding the Lesson:

The teacher can conclude the lesson by creating space and time for students to share their plots and findings in small groups. The opportunity to talk about the plot multiple times to individuals or small groups will offer students an opportunity to refine their thinking, their application of academic language, and learn from each other.

1. Students create a "virtual" poster session in a presentation software that include plots generated from CODAP and key points from their interpretation at their desks. Each partner within the pair prepares to share their data story to their classmates.
2. Partner A stays with their poster and shares their thinking with visiting classmates. Partner B, visits other posters and listens to the presentations.
3. After 10-15 minutes, Partner A and B switch so that Partner B stays with the poster and partner A can move around the room.

Reflections and Additional Recommendations

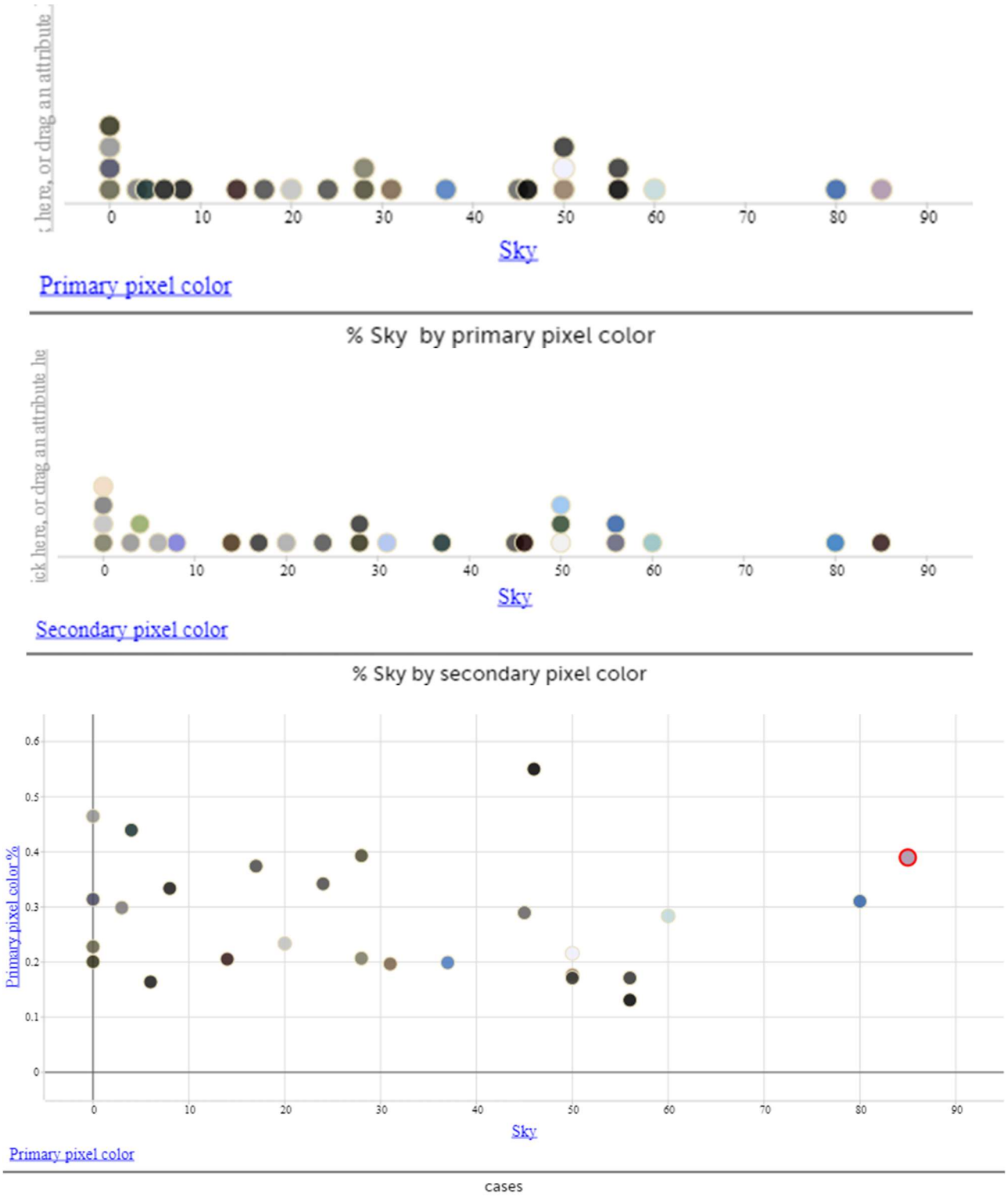
See associated STEW lessons that also have using photographs as data as the starting point, but with different approaches to the investigation.

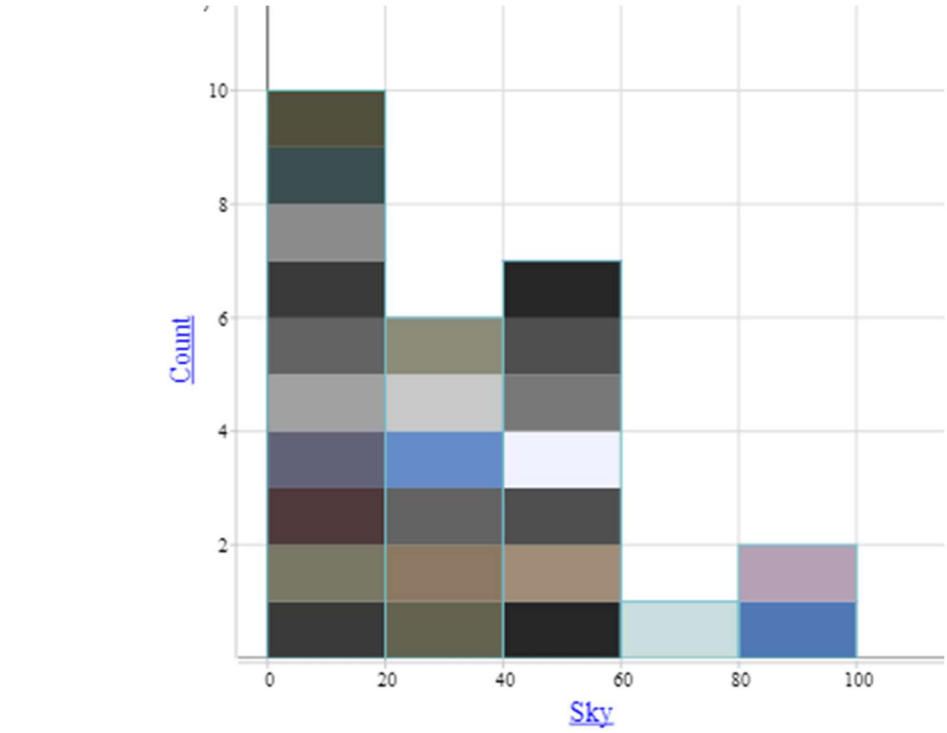
- *Using CODAP to Tell Different Stories*
- *Questioning Throughout the Statistical Problem-Solving Process*

Resources become familiar with CODAP

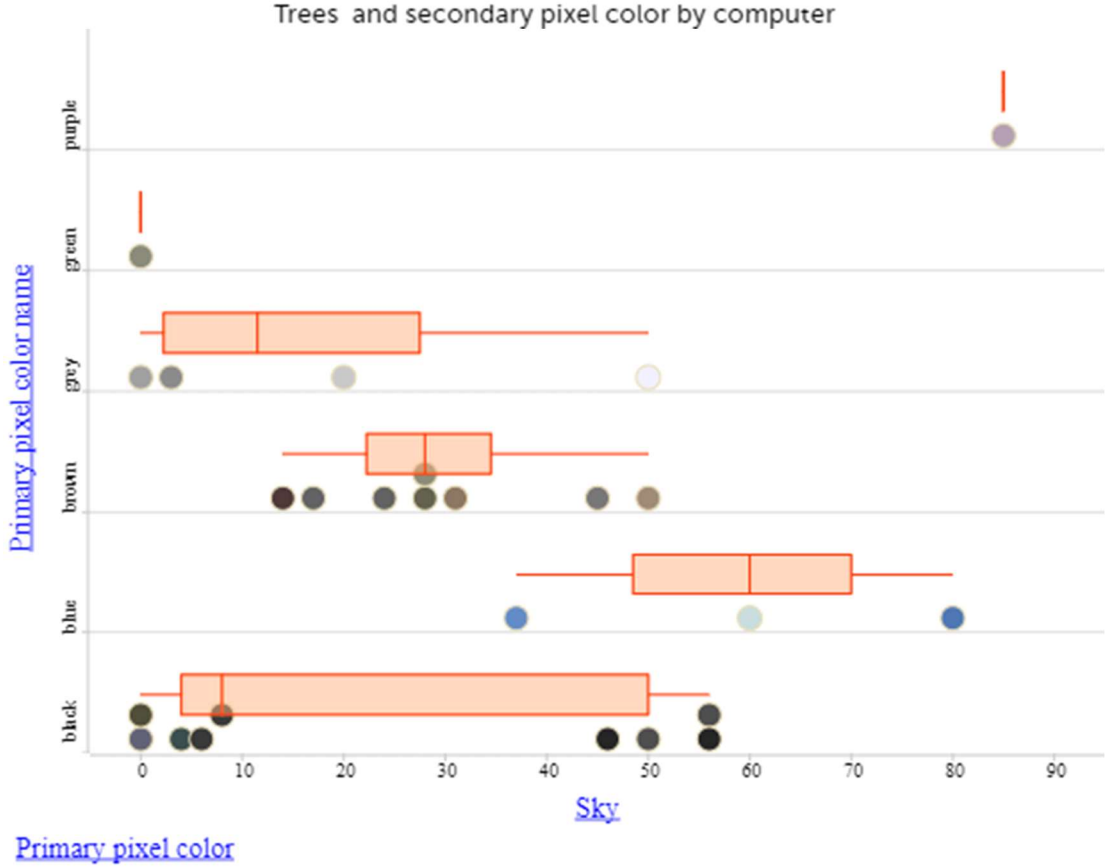
- Karekare Eucation https://karekareeducation.co.nz/category/codap/
- CODAP Getting Started: https://codap.concord.org/help/getting-started

Appendix

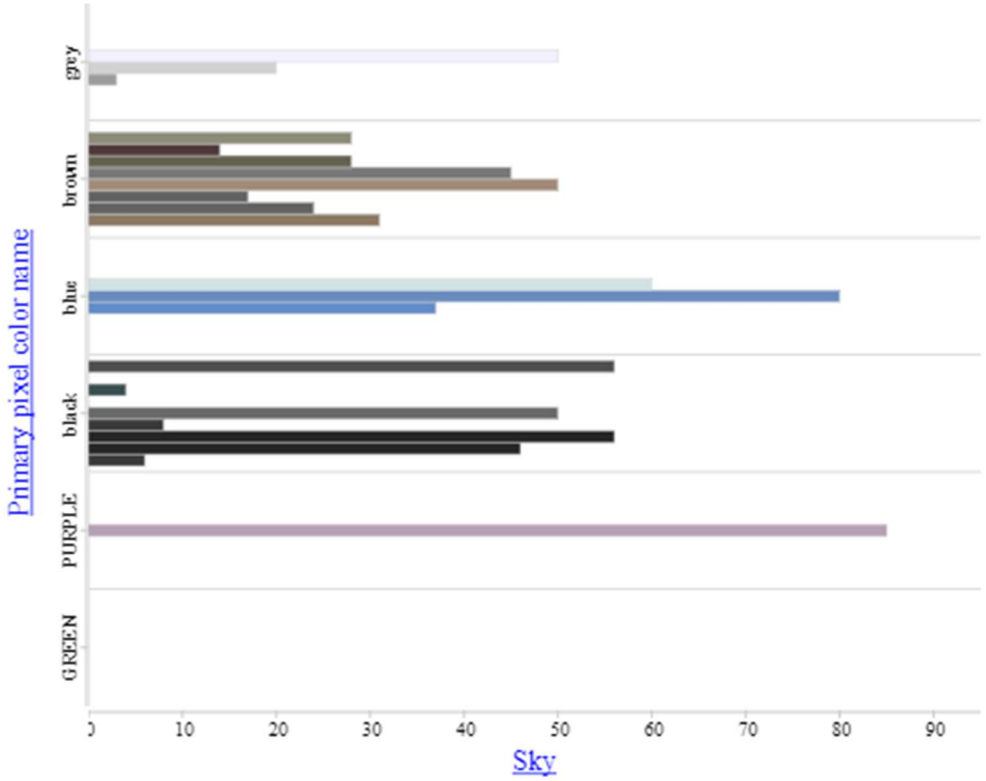

here, or drag an attribute
0
10
20
30
40
50
60
70
80
90
Sky
Primary pixel color
% Sky by primary pixel color
ick here, or drag an attribute he
0
10
20
30
40
50
60
70
80
90
Sky
Secondary pixel color
% Sky by secondary pixel color
0.6
0.5
0.4
0.3
0.2
0.1
0
Primary pixel color %
0
10
20
30
40
50
60
70
80
90
Sky
Primary pixel color
cases

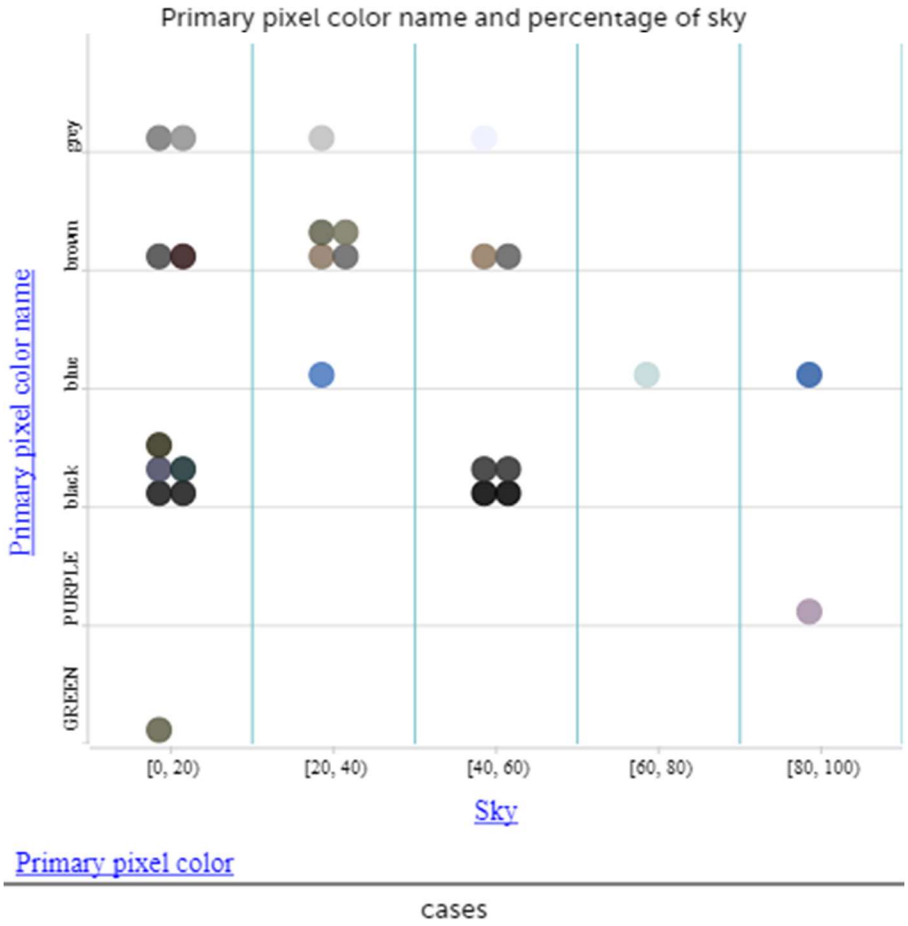


Trees and secondary pixel color by computer



Percentage of sky the photograph including the dominant pixel name


Primary pixel color name
grey
brown
blue
black
PURPLE
GREEN
0
10
20
30
40
50
60
70
80
90
Sky
Primary pixel color


Primary pixel color name and percentage of sky
Primary pixel color name
grey
brown
blue
black
PURPLE
GREEN
[0, 20)
[20, 40)
[40, 60)
[60, 80)
[80, 100)
Sky
Primary pixel color
cases

Additional notes

Statistical investigative questions

From *Questioning Throughout the Statistical Problem-Solving Process*

Background information about posing statistical investigative questions

Arnold's (2013, p. 110–111) research identified six criteria for what makes a good statistical investigative question. The six criteria are:

1. The variable(s) of interest is/are clear and available or can be collected.
2. The group of interest is clear.
3. The intent is clear (e.g., summary, comparison, relationship [association], time series).
4. The [statistical] investigative question can be answered with the data (e.g., question is specific, data can be collected, ethics).
5. The [statistical] investigative question is one that is worth investigating, that it is interesting, that there is a purpose.
6. The [statistical] investigative question allows for analysis to be made of the whole group (e.g., Who is the tallest? is not an investigative question as it is about an individual rather than considering the whole group of interest; What are the heights of the students in our class? considers the whole group of interest – our class).

Types of [statistical] investigative questions - intent (Arnold, 2013)

- Summary investigative questions ask about the overall distribution of the data or what **are** typical and reflects the population/group for which conclusions can be drawn.
- Comparison investigative questions clearly identify the population or groups to be compared. Many comparison situations are considering if one group tends to be bigger, longer, higher than the other group so the investigative questions should include the idea of tendency, e.g., the investigative question includes words or phrases such as tends to, on average, generally.
- Relationship (association) investigative questions can explore paired numerical data. They clearly identify the group for which the paired numerical data has been collected from. The investigative question uses language that signals a relationship (association) is being explored. This might be through a description of the relationship (e.g., if tall people have long arms) or through using the word relationship (e.g., if there is a relationship (association) between height and arm span).
- Time series investigative questions explore patterns over time. Time series situations involve two numerical variables, one of which is time.

For a further discussion on the different types of questions within the statistical problem-solving cycle see Arnold and Franklin (2021) and GAISE II (Bargagliotti et. Al, 2020).

References

Arnold, P. (n.d.). Karekare Education. Retrieved from http://karekareeducation.co.nz/codap/getting-your-students-started-with-codap/

Arnold, P. (2013). *Statistical investigative questions. An enquiry into posing and answering investigative questions from existing data* (Doctoral dissertation, ResearchSpace@ Auckland). https://researchspace.auckland.ac.nz/handle/2292/21305

Arnold, P. & Franklin, C. (2021). What Makes a Good Statistical Question?, Journal of Statistics and Data Science Education, https://www.tandfonline.com/doi/full/10.1080/26939169.2021.1877582

Bargagliotti, A., Franklin, C., Arnold, P., Gould, R., Johnson, S., Perez, L., & Spangler, D. (2020). *Pre-K-12 Guidelines for Assessment and Instruction in Statistics Education (GAISE) report II*. American Statistical Association and National Council of Teachers of Mathematics. https://www.amstat.org/asa/education/Guidelines-for-Assessment-and-Instruction-in-Statistics-Education-Reports.aspx

Common Online Data Analysis Platform [Computer software]. (2014). Concord, MA: The Concord Consortium. https://codap.concord.org/

Gelotte, K. (n.d.). I*mage Color Extract Tool*. Cool PHP Tools. http://www.coolphptools.com/color_extract

Wilkerson, M., & Polman, J. (2020). *Situating Data Science: Exploring How Relationships to Data Shape Learning*. Journal of the Learning Sciences, 29:1, 1-10. DOI: 10.1080/10508406.2019.1705664

Using quadrats - FSC Biology Fieldwork. (n.d.). Field Studies Council. https://www.biology-fieldwork.org/a-level/fieldwork-techniques/vegetation-sampling/using-quadrats/