Unit 9: Data

Lesson 1: Learning from Data
Lesson 2: Exploring One Column
Lesson 3: Filtering and Cleaning Data
Lesson 4: Exploring Two Columns
Lesson 5: Big Data, Crowdsourcing, and Machine Learning
Lesson 6: Machine Learning and Bias
Lesson 7: Project - Tell a Data Story Part 1
Lesson 8: Project - Tell a Data Story Part 2
Lesson 9: Assessment Day

Unit 9 - Lesson 1
Learning from Data

Warm Up

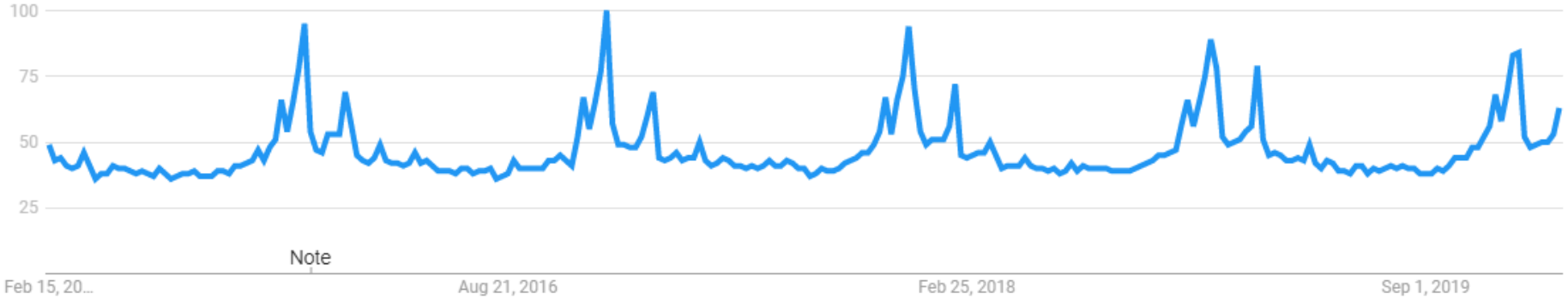
Prompt:

What time of year do people tend to search online for chocolate?

How could you check your guess?

Activity

Google Trends



Searches for chocolate over the past 5 years.

Prompt: What pattern do you notice? What could be the reason for that pattern? Does this data support your earlier guesses?

Two distinctions:

What does the data show?
Why might that be the case?

Let's explore Google Trends together!

https://www.google.com/trends/

What can we do with this tool?

- Add search items (one or more - we can compare graphs)
- Control location, time, search category, and more!
- Look for patterns in the data

Do This:

- With a partner, look for a "data story".
- Try different search items and play with the drop downs to focus on a particular location or time period.
- What interesting things do you uncover?
- What questions can you answer with your charts?
- Define the **what** (facts) and **why** (opinions).

Correlation does not equal Causation

similarities, patterns

this thing caused that thing

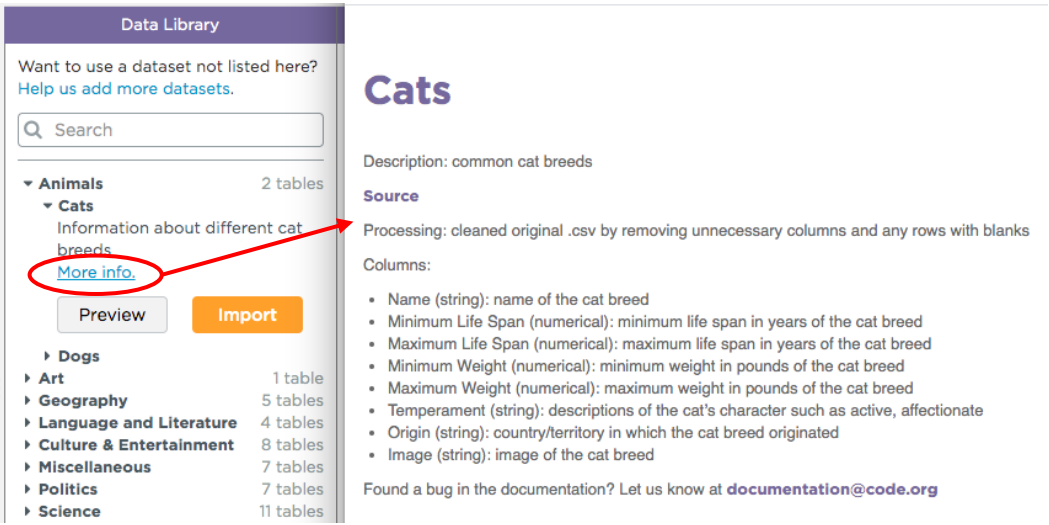
In this unit, we will be making charts to help answer questions:

- *"I think this visualization tells me this..."*
  - Something is more popular than something else
  - Something is more important than something else
  - Something has become more or less searched over time

- *"... but I'm not sure because..."*
  - I don't know exactly how the data was collected
  - This might tell me people searched for green more than red, but it doesn't tell me why they do that or that green is a better color
  - We need more data!

Metadata

data about data



Prompt:

What is the metadata for the chart you created in Google Trends?

Do This:

- Navigate to Level 2
- Open up Data Tab
- Look for the metadata for a table
- Share with a partner:
  - What does this table contain?
  - Where does the data come from?
  - What other useful information is found in the metadata?

Wrap Up

Takeaways:

When looking at visualizations, consider:

- **What** does the data show? - **fact**
- **Why** might that be the case? - **opinion**

Be careful when making assumptions about data:

- **Correlation** does not equal **Causation**

Metadata are data about data:

- It can be changed without impacting the primary data
- Used for finding, organizing, and managing information
- Increases effective use of data by providing extra information
- Allows data to be structured and organized

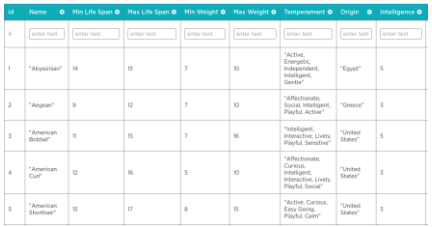
Metadata: data about data

Unit 9 - Lesson 2
Exploring One Column

Warm Up

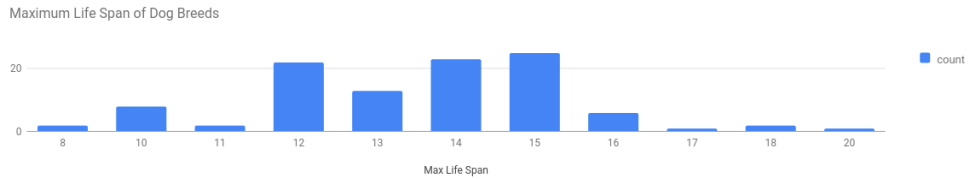
Prompt:

Why do people make visualizations out of data?



id	Name	Min Life Span	Max Life Span	Min Weight	Max Weight	Temperament	Origin	Intelligence
#	enter text	enter text	enter text	enter text	enter text	enter text	enter text	enter text
1	"Abyssinian"	14	15	7	10	"Active, Energetic, Independent, Intelligent, Gentle"	"Egypt"	5
2	"Aegean"	9	12	7	10	"Affectionate, Social, Intelligent, Playful, Active"	"Greece"	3
3	"American Bobtail"	11	15	7	16	"Intelligent, Interactive, Lively, Playful, Sensitive"	"United States"	5
4	"American Curl"	12	16	5	10	"Affectionate, Curious, Intelligent, Interactive, Lively, Playful, Social"	"United States"	3
5	"American Shorthair"	15	17	8	15	"Active, Curious, Easy Going, , Playful, Calm"	"United States"	3

?



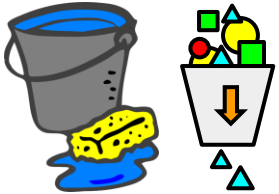
Visualizations can help us:

- Look at lots of data at once
- See patterns that are "invisible" if you just look at the table

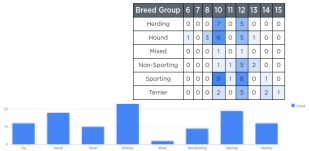
Today we're going to learn how to make two different types of visualizations

The Data Analysis Process

Collect or choose data



Clean and/or filter



Breed Group	6	7	8	10	11	12	13	14	15
Herding	0	0	0	7	0	5	0	0	0
Hound	1	0	3	3	0	5	1	0	0
Mixed	0	0	0	1	0	1	0	0	0
Non-Sporting	0	0	0	1	1	5	2	0	0
Sporting	0	0	0	3	1	8	0	1	0
Terrier	0	0	0	2	0	5	0	2	1

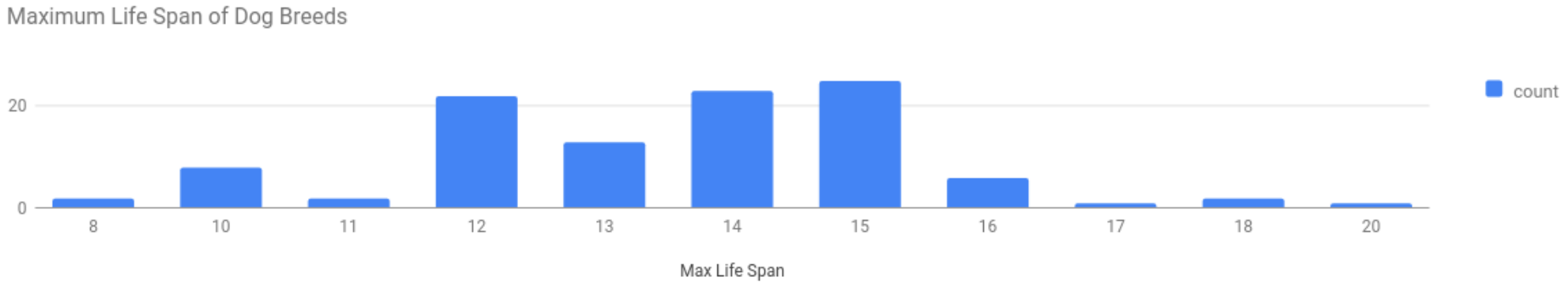
Visualize and find patterns

Generate new Information

For the next three lessons we're going to focus on how data is used to make new information. Afterwards we'll start thinking about how this process is useful in many different contexts.

Activity

Which of these questions does this chart answer?
Be ready to discuss with a classmate

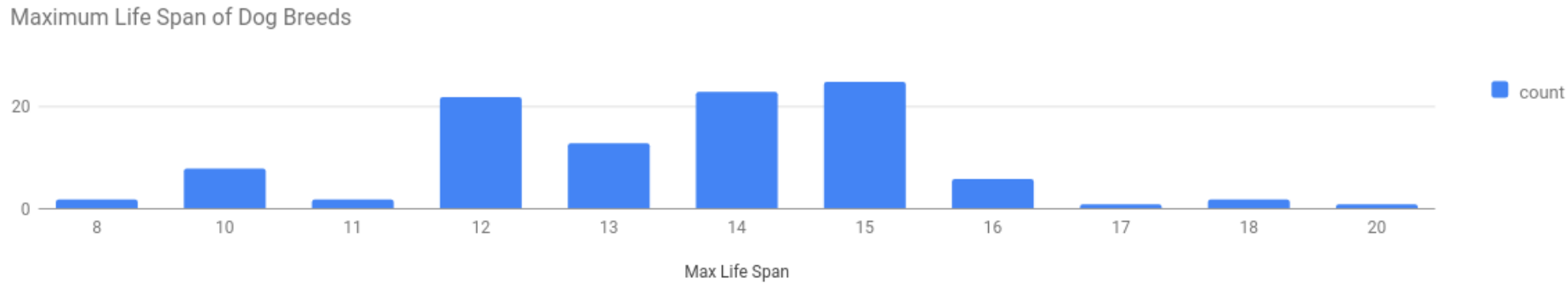


Question 1: What is the most common maximum lifespan of a dog?
Question 2: What is the fluffiest breed of dog?
Question 3: What is the shortest maximum lifespan of a dog?
Question 4: How long will my dog live?
Question 5: What is the longest maximum lifespan of a cat?

Bar Chart: Count how many times each value in the column appears and make a bar at that height.

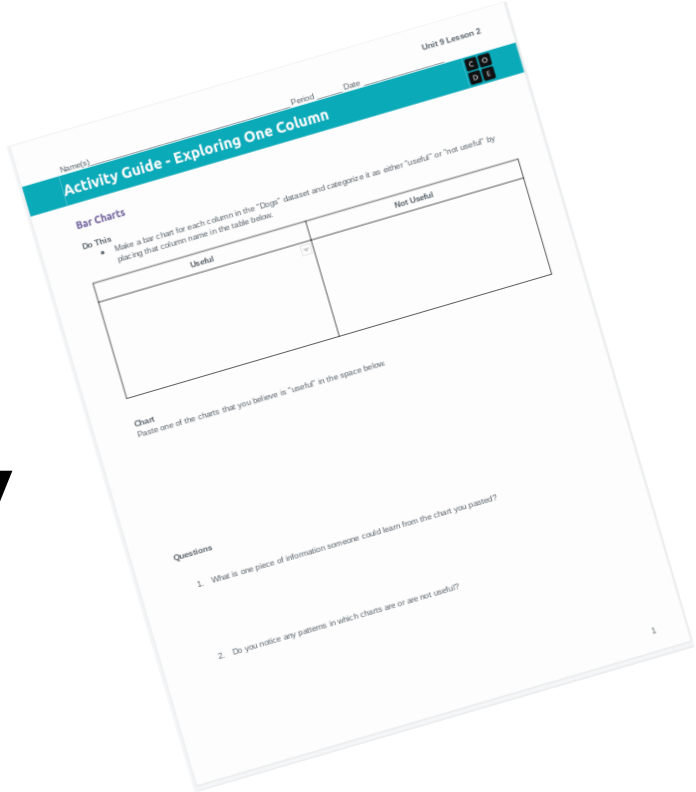
Information we can get out of bar charts:

- What value(s) are most common in this column?
- What value(s) are least common in this column?
- What is the unique list of values in this column?



Let's practice making bar charts ourselves!

You'll need your activity guide, ideally in digital form.



Unit 9 Lesson 2

Name(s) Period Date

Activity Guide - Exploring One Column

Bar Charts

Do This

- Make a bar chart for each column in the "Dogs" dataset and categorize it as either "useful" or "not useful" by placing that column name in the table below.

Useful	Not Useful

Chart

Paste one of the charts that you believe is "useful" in the space below.

Questions

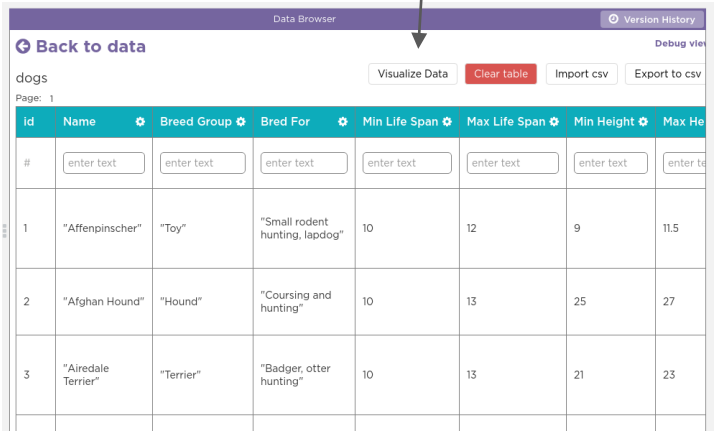
1. What is one piece of information someone could learn from the chart you pasted?
2. Do you notice any patterns in which charts are or are not useful?

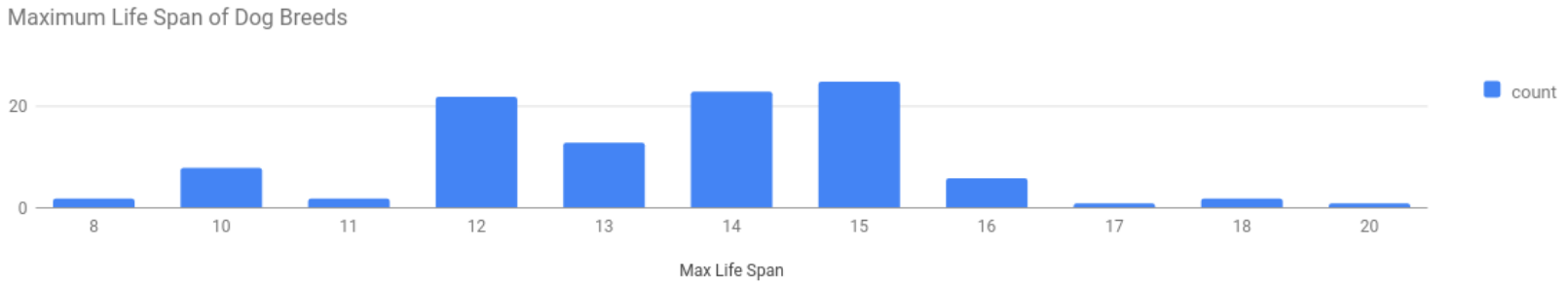
1

Do This: Make the Dogs Bar Chart

- Go to Lesson 2, Level 2
- Open the Dogs table in the Data Tab
- Open the Data Visualizer
- Set the following values
  - Chart Type: "Bar Chart"
  - Values: "Maximum Life Span"
- Once you've made the chart, help others around you

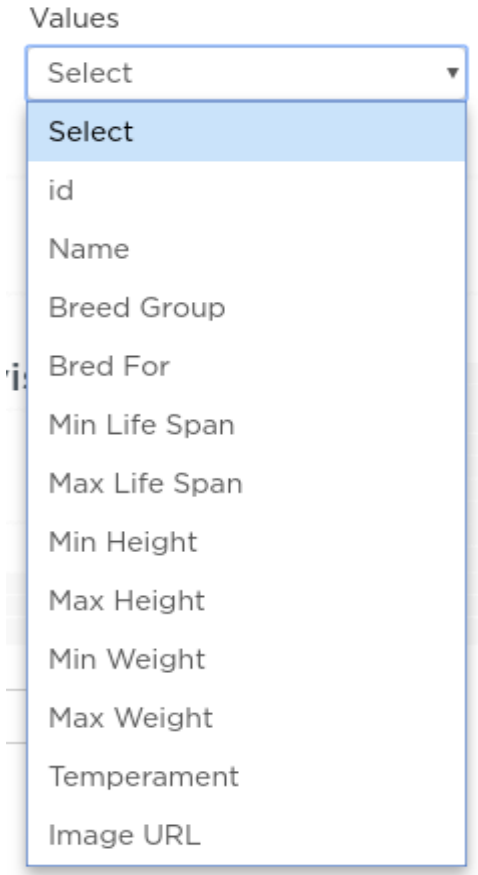
Click "Visualize Data"





Do This: Fill out page 1 of your activity guide

- Make a bar chart for every column in the Dogs table and decide if the chart is "useful" or "not useful"
- Paste one of the charts you think is "useful" into your activity guide
- Answer the questions at the bottom

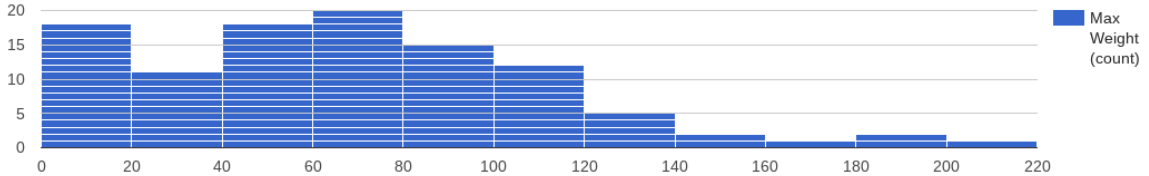


This bar chart of "Temperament" isn't very useful because all of the values are unique. In general if your column has too many unique values it gets difficult to make any sense of them or find patterns.

There are cases where your column has many different values where a different kind of chart is useful. Let's go look at an example.

This is the bar chart for "Max Weight", it's a little difficult to read

In the Visualizer make a "Histogram" for "Max Weight" with a "Bucket Size" of 20



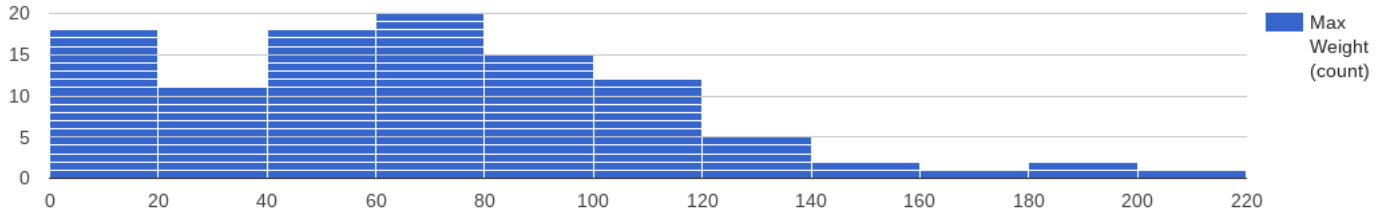
Prompt: Which of these makes it easier to understand the data?
What do you think the "histogram" is doing to visualize the data differently?

Histogram: Similar to a bar chart, but first all numbers in a range or "bucket" are grouped together. For example, the chart below has a bucket size of 20 so the numbers 41, 48, and 53 would all be placed in the same bucket between 40 and 60.

Histograms can only be created with numeric data but can be useful when a normal bar chart may be difficult to read.

Information we can get out of histograms:

- What range of value(s) are most common in this column?
- What range value(s) are least common in this column?
- What ranges of values do or do not appear in this column?



Do This: Fill out page 2 of your activity guide

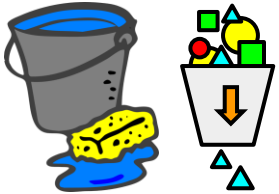
- Try out different bucket sizes for the dogs "max weight" and decide which one you think is most helpful for reading the chart.

- Make a histogram for another column in the "Dogs" table and decide on a range size that helps you find an interesting pattern.

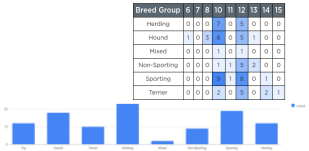
Wrap Up

The Data Analysis Process

Collect or Choose Data



Clean and/or Filter



Visualize and Find Patterns

New Information

Prompt: Which steps of the Data Analysis Process did you see in today's activity? Where did you see them?

Hint: There's one step we didn't do at all!

Key Takeaways

- Programs (like the Data Visualizer) can help process data so we can understand it and learn.
- Charts and other visualizations can help both find and communicate what we've learned from data
- Bar charts and histograms are two common chart types for exploring one column of data in a table.

Unit 9 - Lesson 3
Filtering and Cleaning Data

Warm Up

Imagine you have used a survey to collect information from students.

Collect or Choose Data

All of that data is now stored in a table. You are excited to dig into the data and see what you can learn. Let's go!

???

Activity

Do This:

- Navigate to Level 2 on Code Studio
- Open the data tab
- Familiarize yourself with the imported table
- Open the Data Visualizer
- Make charts:
  - Average Hours of Sleep
  - Favorite Subject

Prompt:

Discuss your charts with a partner.

What problems came up when trying to create these charts?

What problems do you see in the data?

Cleaning Data

When does data need to be cleaned?

- Data is incomplete
- Data is invalid
- Multiple tables are combined into one

What leads to "messy" data?

- Users enter in different types of data ("two", 2)
- Users use different abbreviations to represent the same information ("February", "Feb", "Febr")
- Data may have different spellings ("color", "colour") or inconsistent capitalization ("spring", "Spring")

Cleaning Data

Goal: To clean data without changing meaning

Method:

- Look through the data manually. Find and fix messy data.
- Use a program to find and fix messy data.

We will use the manual method today.

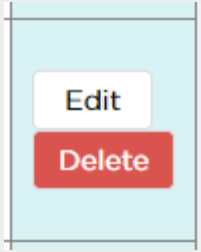
Do This:

With a partner:

- Clean the Student Info table
- Look for:
  - Different types of data ("two", 2)
  - Different abbreviations to represent the same information ("February", "Feb", "Febr")
  - Different spellings ("color", "colour")
  - Inconsistent capitalization ("spring", "Spring")
- Manually update cells with messy data so they are consistent with other cells, while not changing the meaning of the data.

* Note: Not all of these are present in this dataset

To edit a cell, click on the "edit" button at the end of the row. Click "save" after you've finished.



Do This:

- Make charts:
  - Average amount of sleep
  - Favorite class

Discuss:

- Do your charts match others in the class? Why or why not?

Prompt:

What if I only want to look at a subset of my data? How could I do this?

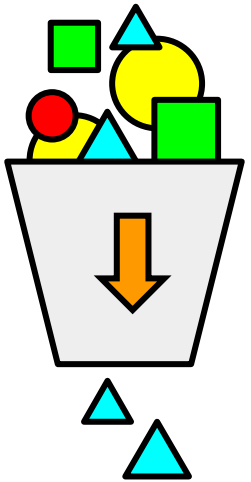
For example: I only want to investigate dogs with low shedding

Filtering Data

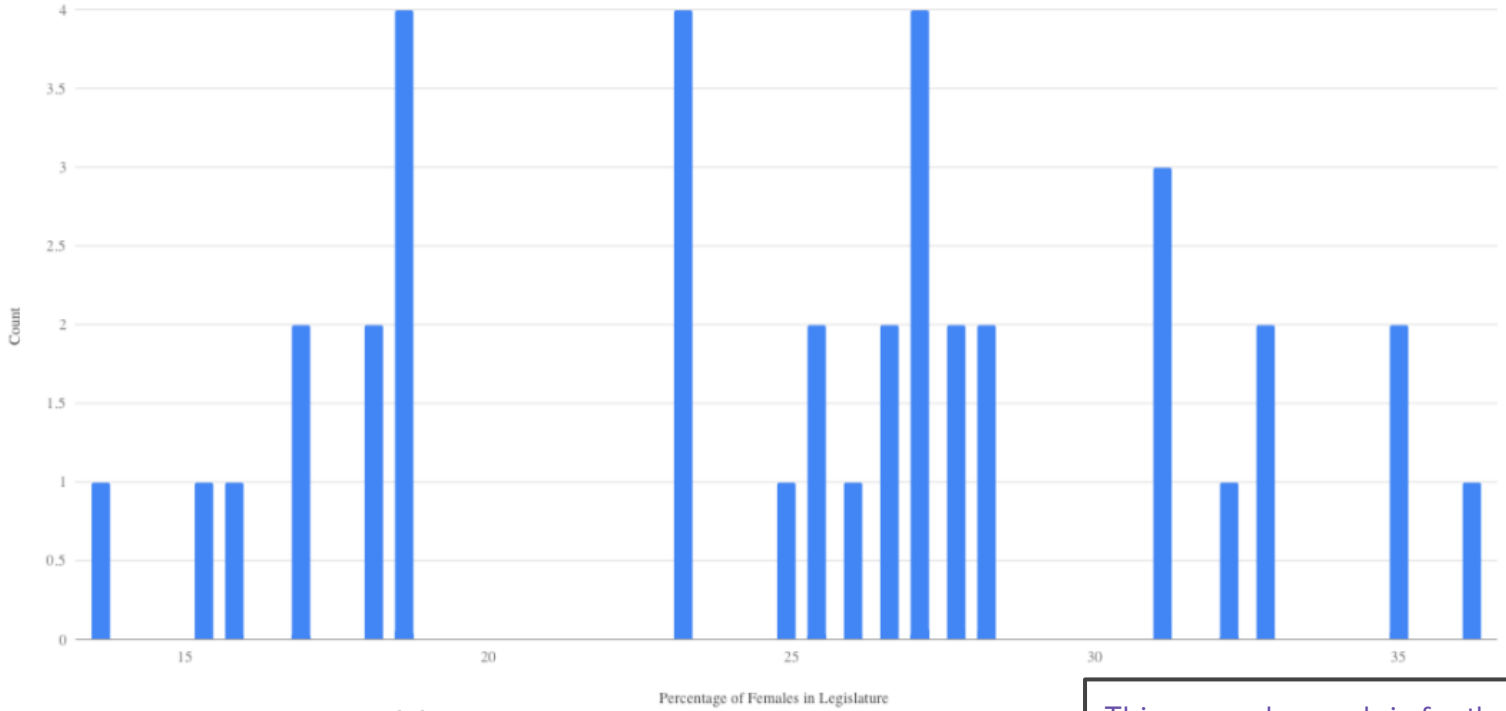
Filtering data allows the user to look at a subset of the data.

In Unit 5, we filtered data programmatically using traversals to gain insight into knowledge from data.

Software programs with built in tools (like the Data Visualizer) can also be used to filter data.



Demo:



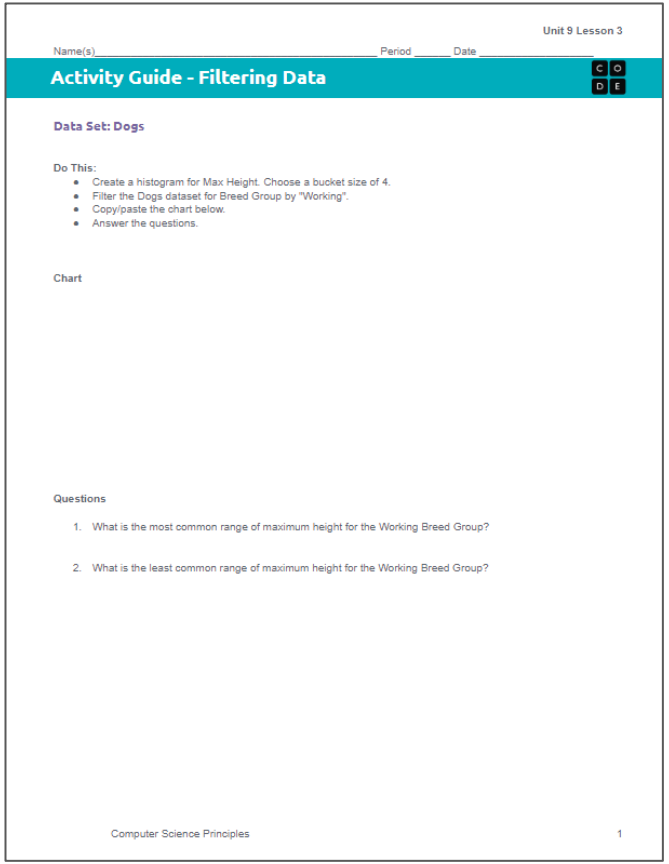
Created using Code.org App Lab on 2020/11/02 from the Female State Legislators table in project

Values: Percentage of Females in Legislature, Filtered State to "Illinois"

This example graph is for the state of Illinois, your state's graph may look different.

Do This:

- Navigate to Level 3
- Work through the Activity Guide
- Copy/Paste the charts you make into the Activity Guide



Unit 9 Lesson 3

Name(s)_____ Period _____ Date _____

Activity Guide - Filtering Data

Data Set: Dogs

Do This:

- Create a histogram for Max Height. Choose a bucket size of 4.
- Filter the Dogs dataset for Breed Group by "Working".
- Copy/paste the chart below.
- Answer the questions.

Chart

Questions

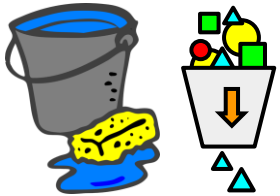
1. What is the most common range of maximum height for the Working Breed Group?
2. What is the least common range of maximum height for the Working Breed Group?

Computer Science Principles 1

Wrap Up

The Data Analysis Process

Collect or Choose Data



Clean and/or Filter

Visualize and Find Patterns

New Information

Prompt: Why is "Clean and/or Filter" an important part of the Data Analysis Process?

What are situations when you would filter vs. clean your data?

Unit 9 - Lesson 4
Exploring Two Columns

Warm Up

Prompt:

Imagine you wanted to know which hour of the day you and your classmates are happiest.

What kind of data would you collect?

How do you think you'd analyze it?

Bar charts and histograms are only useful for looking at one column of data.

If we want to look at relationships between two pieces of information (like time of day and happiness) we'll need ways to visualize data that look at two columns of data at the same time.

Activity

Breed Group	Max Life Span
"Working"	12
"Toy"	12
"Working"	12
"Mixed"	13
"Working"	10
"Non-sporting"	15
"Mixed"	13

Maximum Life Span vs Breeding Group of Dog Breeds

Breed Group	8	10	11	12	13	14	15	16	17	18	20
Herding	0	0	0	4	2	3	2	1	0	0	0
Hound	1	2	0	4	3	4	3	1	0	0	0
Mixed	0	0	0	0	2	0	0	0	0	0	0
Non-Sporting	0	0	0	0	2	1	5	1	0	0	0
Sporting	0	0	1	5	1	6	5	1	0	0	0
Terrier	0	0	0	0	1	2	5	1	0	0	1
Toy	0	0	0	1	0	4	3	1	1	2	0
Working	1	6	1	8	2	3	2	0	0	0	0

A **crosstab chart** counts how many times combinations of values appear. Arrows show where that row in the data table would be counted in the chart.

Prompt:

- How many "Herding" breeds live a maximum of 12 years?
- What is the most common maximum life span for "Working" breeds?
- Which breed group lives the shortest?
- Which breed group lives the longest?
- How do you know? How confident are you in your answers?

Cross Tab: Counts how often pairs of values in two columns appear.

Useful for:

- Finding the most / least common combinations of values in two columns
- Finding patterns across two columns
- Exploring two columns when one or both are strings.

Not useful:

- If either column has too many values (the chart would be enormous)

Breed Group	8	10	11	12	13	14	15	16	17	18	20
Herding	0	0	0	4	2	3	2	1	0	0	0
Hound	1	2	0	4	3	4	3	1	0	0	0
Mixed	0	0	0	0	2	0	0	0	0	0	0
Non-Sporting	0	0	0	0	2	1	5	1	0	0	0
Sporting	0	0	1	5	1	6	5	1	0	0	0
Terrier	0	0	0	0	1	2	5	1	0	0	1
Toy	0	0	0	1	0	4	3	1	1	2	0
Working	1	6	1	8	2	3	2	0	0	0	0

Let's practice making more crosstabs ourselves!

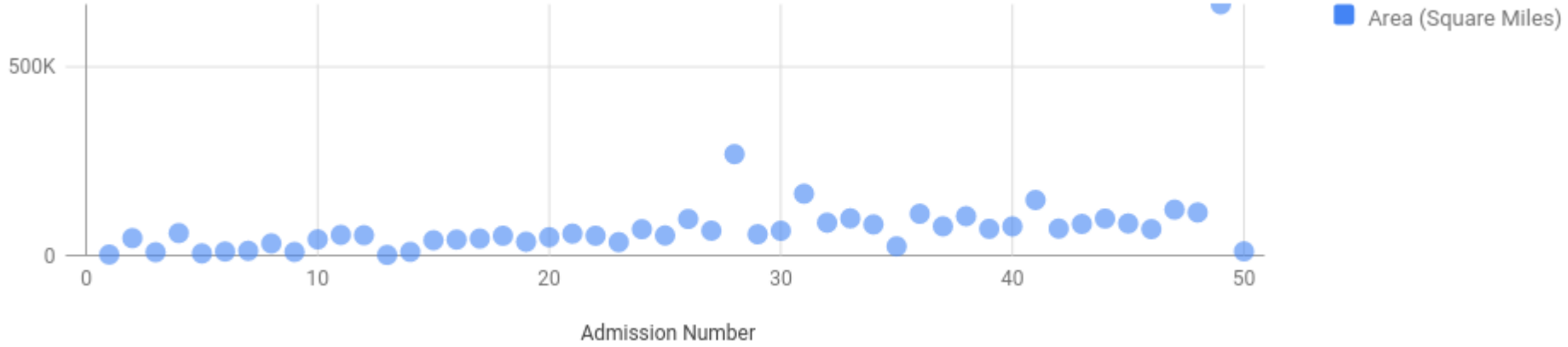
You'll need your activity guide, ideally in digital form.

Do This: Fill out page 1 of your activity guide

- Make the two different crosstab charts indicated on this sheet
- Answer the questions for each chart

This table has a dot for each state. It shows the size of the states in square miles vs. the order it was added to the USA.

Prompt: Is there a pattern? How can you tell?



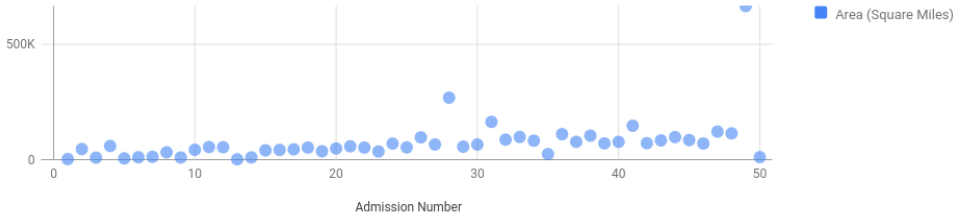
Scatter: Shows combinations of values from two columns

Useful for:

- Seeing patterns and trends between two values
- Numeric data with lots of different values

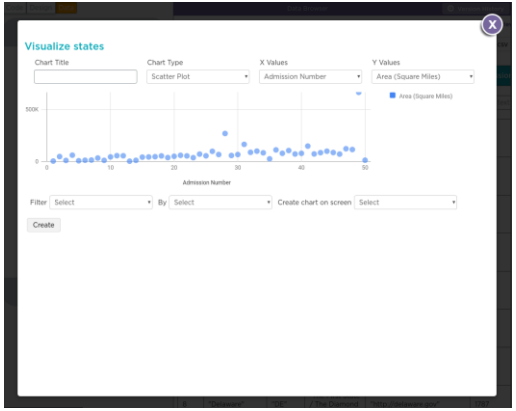
Not useful:

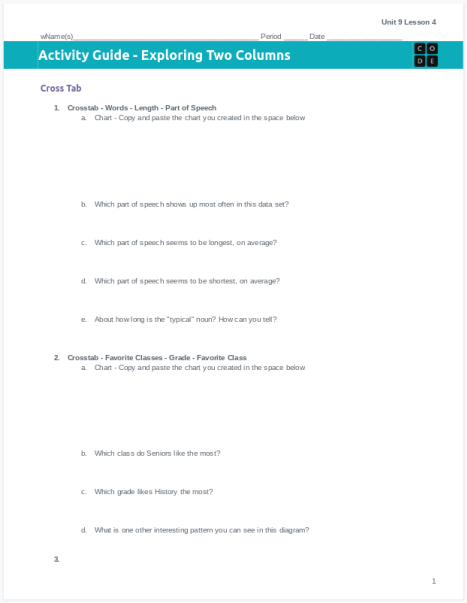
- Lots of repeated values



Do This: Complete both sides of your activity guide

- On Code Studio to Lesson 4 Level 2
- Practice making crosstab and scatter charts to answer different questions. Fill in the information you collect in activity guide.





Unit 9 Lesson 4

wName(s)_____ Period____ Date_____

Activity Guide - Exploring Two Columns

Cross Tab

1. Crosstab - Words - Length - Part of Speech
   a. Chart - Copy and paste the chart you created in the space below
   b. Which part of speech shows up most often in this data set?
   c. Which part of speech seems to be longest, on average?
   d. Which part of speech seems to be shortest, on average?
   e. About how long is the "typical" noun? How can you tell?
2. Crosstab - Favorite Classes - Grade - Favorite Class
   a. Chart - Copy and paste the chart you created in the space below
   b. Which class do Seniors like the most?
   c. Which grade likes History the most?
   d. What is one other interesting pattern you can see in this diagram?
3.

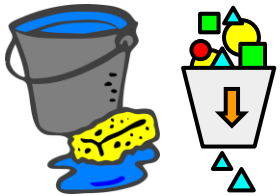
1

Wrap Up

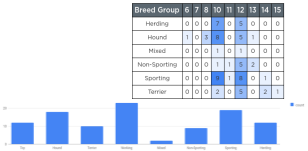
We now know a lot of different ways to Visualize and Find Patterns

The Data Analysis Process

Collect or Choose Data



Clean and/or Filter



Visualize and Find Patterns

New Information

This side is good for knowing what's in your dataset at all. What values are present, what's their range, where are the clusters of values, and so on.

This side is good for finding relationships and patterns across different columns. You might start being able to see how changes to one value impact another.

How many columns?

1 column

2 columns

How many values?

Few

Many

Are the values strings?

Yes

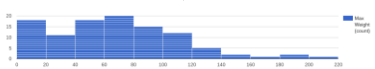
No

How many values?

Few

Many

Bar chart



Histogram

Breed Group	6	7	8	10	11	12	13	14	15
Herding	0	0	0	7	0	5	0	0	0
Hound	1	0	3	8	0	5	1	0	0
Mixed	0	0	0	1	0	1	0	0	0
Non-Sporting	0	0	0	1	1	5	2	0	0
Sporting	0	0	0	9	1	8	0	1	0
Terrier	0	0	0	2	0	5	0	2	1
Toy	0	0	0	4	0	5	1	1	1
Working	1	2	5	10	1	4	0	0	0

Crosstab

Scatter Plot

Key Takeaways

- We can develop insights and knowledge about our world from manipulating and visualizing data, in particular by finding patterns

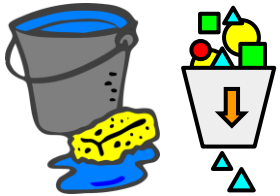
- When investigating two columns of data we can observe patterns different values move together (are correlated). We cannot know for certain the cause of the correlation.

Unit 9 - Lesson 5
Big, Open, and Crowdsourced Data

Warm Up

The Data Analysis Process

Collect or Choose Data



Clean and/or Filter

Breed Group	6	7	8	10	11	12	13	14	15
Herding	0	0	0	7	0	5	0	0	0
Hound	1	0	3	8	0	5	1	0	0
Mixed	0	0	0	1	0	1	0	0	0
Non-Sporting	0	0	0	1	1	5	2	0	0
Sporting	0	0	0	9	1	8	0	1	0
Terrier	0	0	0	2	0	5	0	2	1

Visualize and Find Patterns

New Information

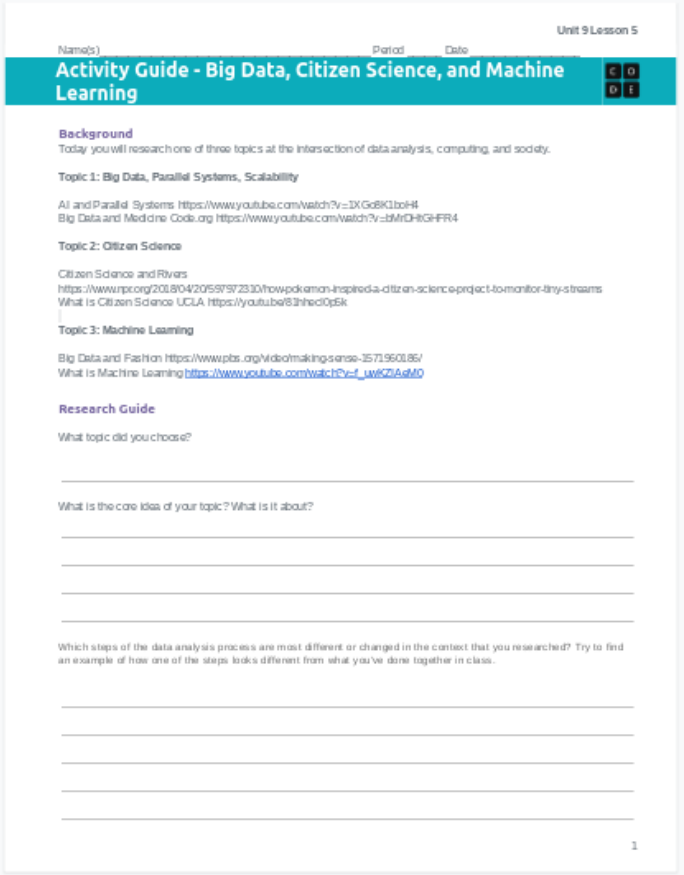
Prompt: With a partner review the data analysis process and for each step talk through:

- What is this step and why is it important?
- Where have we done this step together?
- What could go wrong if you do this step poorly.

Activity

Prompt: With a partner

- Choose one of the topics
- Watch the related videos / listen to the podcasts
- Take notes and be ready to share responses to the three questions below



Unit 9 Lesson 5

Name(s) _____ Period ____ Date _____

Activity Guide - Big Data, Citizen Science, and Machine Learning

Background

Today you will research one of three topics at the intersection of data analysis, computing, and society.

Topic 1: Big Data, Parallel Systems, Scalability

AI and Parallel Systems https://www.youtube.com/watch?v=1XGo8K1boH4
Big Data and Medicine Code.org https://www.youtube.com/watch?v=bMrDHtGHFR4

Topic 2: Citizen Science

Citizen Science and Rivers
https://www.npr.org/2018/04/20/599797330/how-pokemon-inspired-a-citizen-science-project-to-monitor-tiny-streams
What is Citizen Science UCLA https://youtu.be/83hhecl0pEk

Topic 3: Machine Learning

Big Data and Fashion https://www.pbs.org/video/making-sense-1571960186/
What is Machine Learning https://www.youtube.com/watch?v=_uwKZIAeM0

Research Guide

What topic did you choose?

What is the core idea of your topic? What is it about?

Which steps of the data analysis process are most different or changed in the context that you researched? Try to find an example of how one of the steps looks different from what you've done together in class.

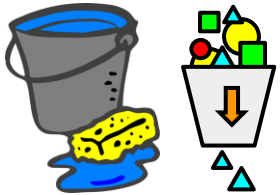
1

Discuss: Multiple members of each group should present on

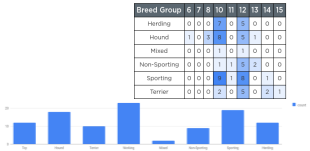
- What the topic is
- The key vocabulary you were responsible for
- How this concept uses or modifies the data analysis process
- Examples of the problems this technique is being used to solve

The Data Analysis Process

Collect or Choose Data



Clean and/or Filter



Visualize and Find Patterns

New Information

Wrap Up

Open Data

- "sharing data with others so they can can analyze it"
- Open data is publicly available data shared by governments, organizations, and others
- Making data open help spread useful knowledge or creates opportunities for others to use it to solve problems

Citizen Science and Crowdsourcing

- "collecting data from others so you can analyze it"
- Crowdsourcing is the practice of obtaining input or information from a large number of people via the Internet.
- Citizen science is research where some of the data collection is done by members of the public using own computing devices which leads to solving scientific problems
- Crowdsourcing offers new models for collaboration, such as connecting businesses or social causes with funding
- Both are examples of how human capabilities can be enhanced by collaboration via computing

Big data

- "Collect huge amounts of data so we can learn even more from it"
- The size of the datasets we analyzed impacts how much information can be extracted
- As a result, in business, science, and many other contexts people are working with increasingly big data sets
- When data gets too big it can no longer be processed on one computer. Cloud computing or parallel systems are sometimes used to help process all that information.
- In general scalability of your system is important to consider when working with big data. You want your system to be able to work even as you're using more and more data.

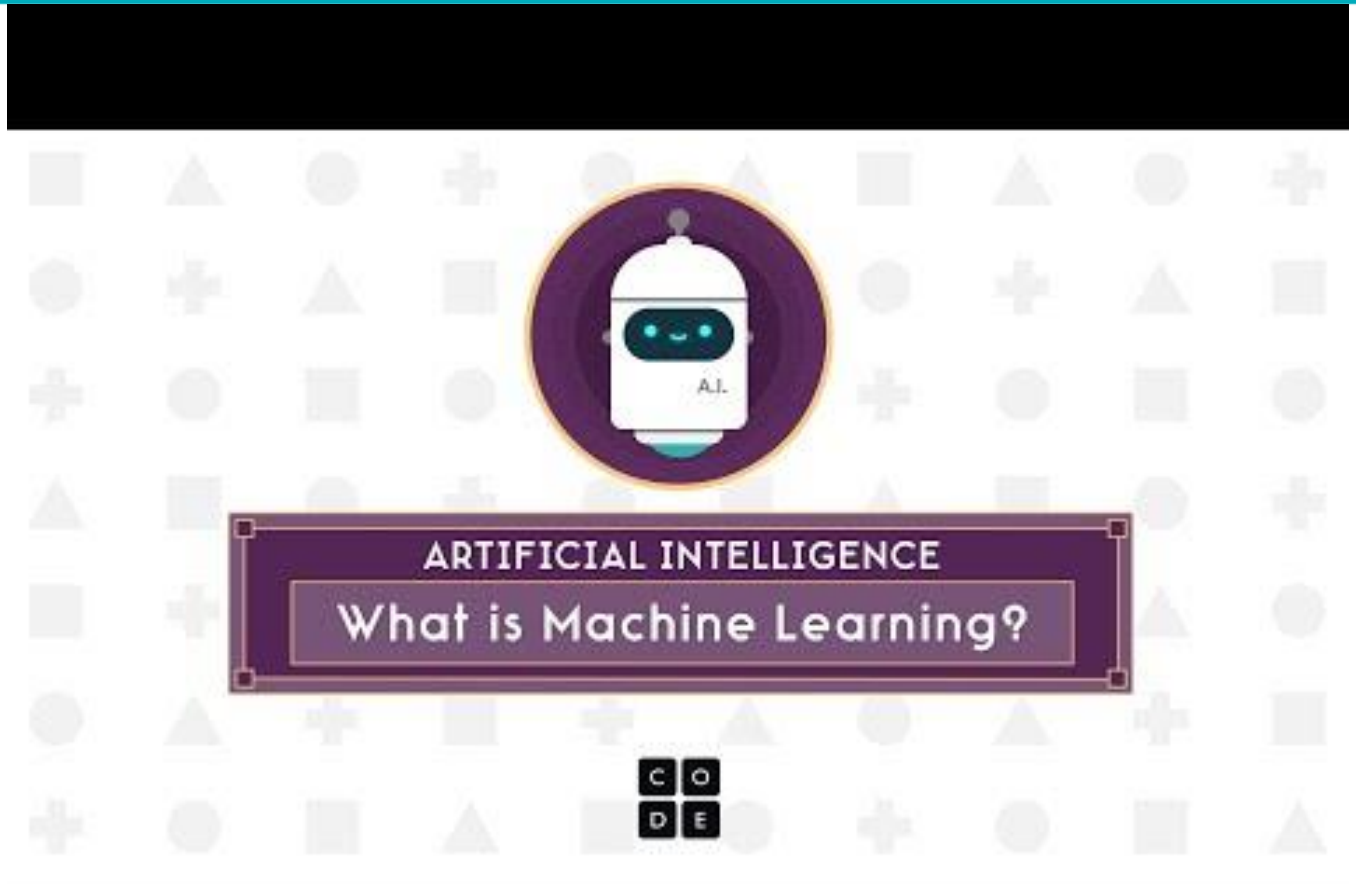
Unit 9 - Lesson 6
Machine Learning and Bias

Warm Up

Prompt:

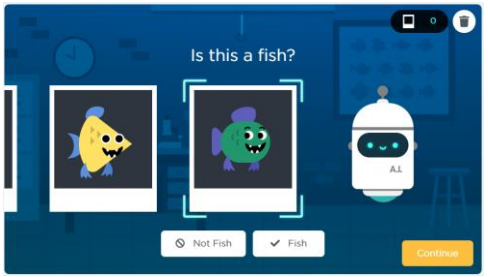
How can machines "learn"?

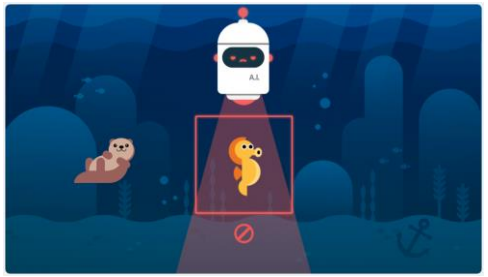
Activity

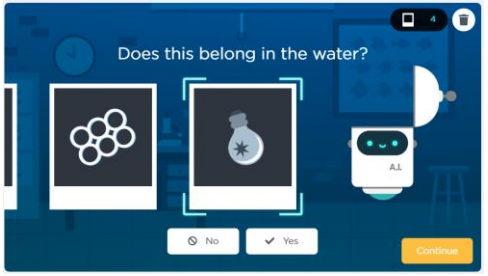

A.I.
ARTIFICIAL INTELLIGENCE
What is Machine Learning?
C O
D E

Do This: Complete Levels 3-5

Consider: What potential problems do you think come up when creating training sets?







Prompts:

How well did A.I. do?

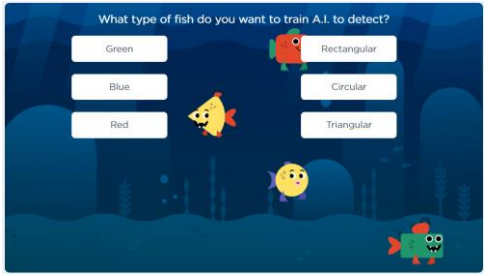
How do you think it decided what to include in the ocean?

A.I.
ARTIFICIAL INTELLIGENCE
Training Data & Bias
C O
D E

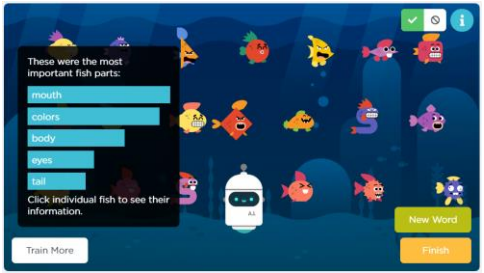
Prompts:

How do you think your training data influenced the results that A.I. produced?

Do This: Complete Levels 7-8



Consider: Is it fair to use AI to judge something by looks? What unintended bias could this cause?



Prompts:

How could biased data result in problems for artificial intelligence? What are ways to address this?

Prompts:

How can computing innovations which make use of Machine Learning reflect existing human bias?

How could it be used to discriminate against groups of individuals?

How can that bias be minimized?

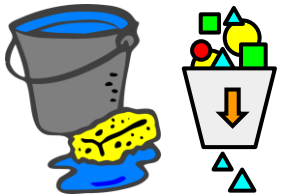
A.I.
ARTIFICIAL INTELLIGENCE
Impact on Society
C O
D E

Wrap Up

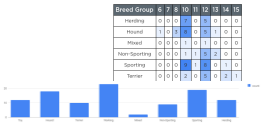
Prompt: Which steps of this process do you think have to be done by humans? Would you be concerned if any of them were automated?

The Data Analysis Process

Collect or Choose Data



Clean and/or Filter



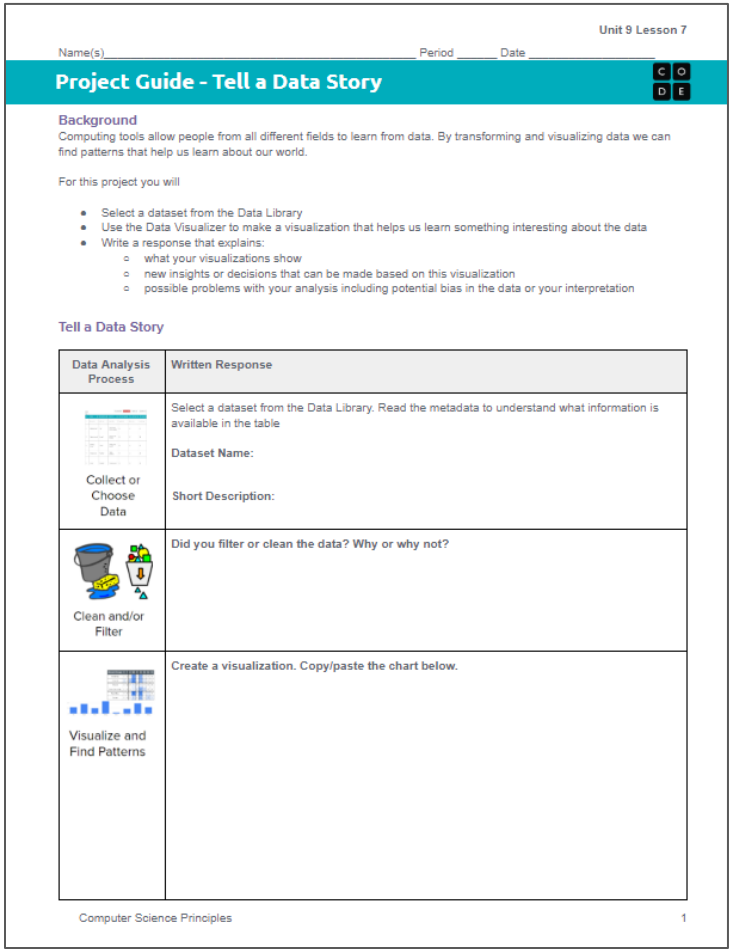
Visualize and Find Patterns

New Information

Unit 9 - Lesson 7
Project - Tell a Data Story Part 1

Warm Up

Activity



Unit 9 Lesson 7

Name(s)_____ Period _____ Date _____

Project Guide - Tell a Data Story

Background

Computing tools allow people from all different fields to learn from data. By transforming and visualizing data we can find patterns that help us learn about our world.

For this project you will

- Select a dataset from the Data Library
- Use the Data Visualizer to make a visualization that helps us learn something interesting about the data
- Write a response that explains:
  - what your visualizations show
  - new insights or decisions that can be made based on this visualization
  - possible problems with your analysis including potential bias in the data or your interpretation

Tell a Data Story

Data Analysis Process	Written Response
Collect or Choose Data	Select a dataset from the Data Library. Read the metadata to understand what information is available in the table **Dataset Name:** **Short Description:**
Clean and/or Filter	**Did you filter or clean the data? Why or why not?**
Visualize and Find Patterns	**Create a visualization. Copy/paste the chart below.**

Computer Science Principles 1

Project - Tell a Data Story

Do This: Today you are working on Page 1 of the Project Guide that covers these steps of the Data Analysis Process:

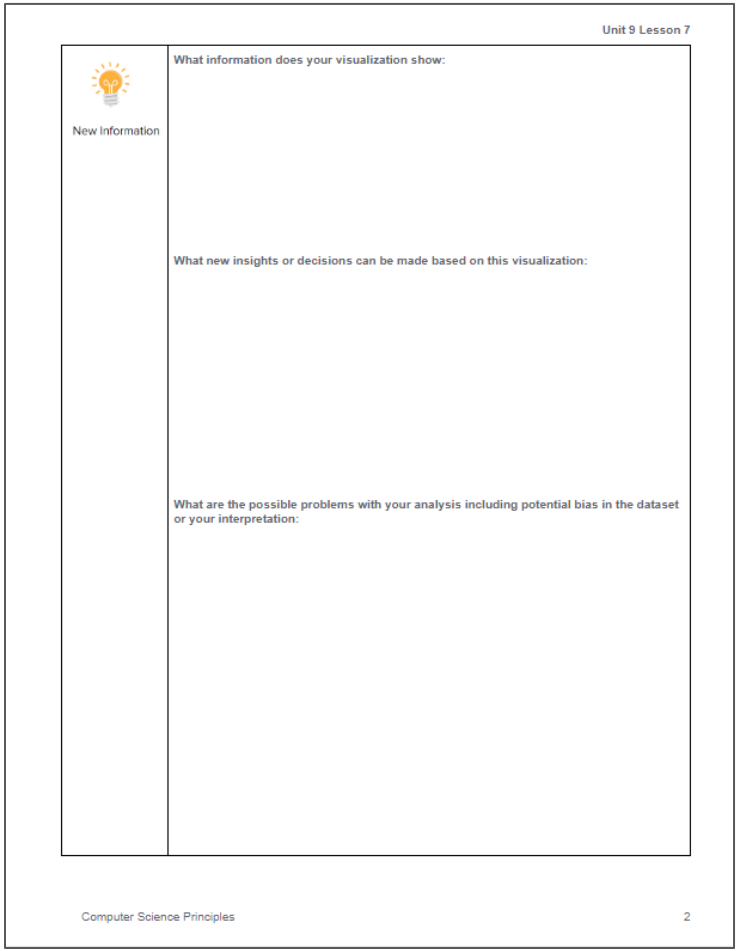
- Choose or Collect Data
- Clean and/or filter
- Visualize and Find Patterns

Wrap Up

Unit 9 - Lesson 8
Project - Tell a Data Story Part 2

Warm Up

Activity



Unit 9 Lesson 7

New Information

What information does your visualization show:

What new insights or decisions can be made based on this visualization:

What are the possible problems with your analysis including potential bias in the dataset or your interpretation:

Computer Science Principles 2

Project - Tell a Data Story

Do This: Today you are working on Page 2 of the Project Guide that covers this step of the Data Analysis Process:

- New Information

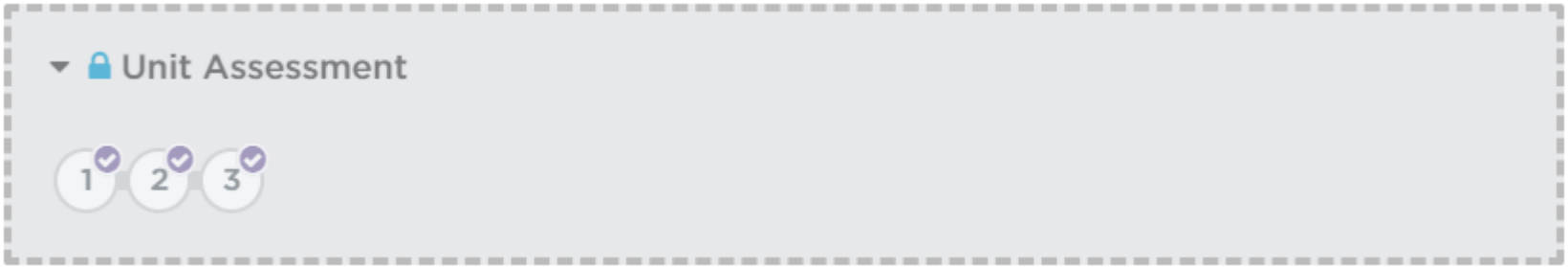
Wrap Up

Reminder:

Before turning in your completed Project Guide, check the rubric on Page 3.

Category	Extensive Evidence	Convincing Evidence	Limited Evidence	No Evidence
Collect or Choose Data	Dataset is correctly identified and description is complete.	Dataset is correctly identified and description is mostly complete.	Dataset is correctly identified and description is somewhat complete.	Dataset is not identified or description is missing.
Clean/Filter Data	Explanation for cleaning and/or filtering is complete and reasonable.	Explanation for cleaning and/or filtering is complete and mostly reasonable.	Explanation for cleaning and/or filtering is somewhat complete or somewhat reasonable.	Explanation for cleaning and/or filtering is incomplete.
Visualize and Find Patterns	Visualization is readable and includes a title, and citation.	Visualization is mostly readable and includes a title, and citation.	Visualization is somewhat readable and/or is missing a title or citation. .	Visualization is unreadable or missing.
New Information: Interpreting the visualization	The visualization is thoroughly explained.	The visualization is mostly explained.	The visualization is somewhat explained.	The visualization is not explained.
New Information: Insights or Decisions	Insights or decisions are reasonable and effectively linked to information displayed in the visualization	Insights or decisions are mostly reasonable and effectively linked to information displayed in the visualization	Insights or decisions are mostly reasonable and somewhat effectively linked to information displayed in the visualization	Insights or decisions are missing.
New Information: Bias	Possible problems with analysis or potential bias are reasonable and thoughtfully addressed.	Possible problems with analysis or potential bias are mostly reasonable and thoughtfully addressed.	Possible problems with analysis or potential bias are somewhat reasonable and addressed.	Possible problems with analysis or potential bias are not addressed.

Unit 9 - Lesson 9
Assessment Day

Activity

Unit Assessment