Intro to Data Science in p5.js

Adam Prado

Goals of Unit (CS, Stats, Civics)

- Further developing p5 coding skills.
- Load, access, and manipulate date with p5.
- Practice with visualizing data.
- Making conclusions based on examining data.
- Checking for errors and bias in data or conclusions.
- Create something to improve yourself or your community.
- Brief exposure to “Data Science”

Unit Overview

1. Intro to Data Science + importing data into p5.

2. Creating functions with univariate data

* Cleaning and filtering data sets

3. Bar Graph (map ())

4. Circle Graph

5. Bivariate Data + Line Graph

6. Scatter Plot

7. Linear regression (causation v. correlation

8. Analyzing results and making conclusions

*. Big Data, Machine Learning,

9. Concerns about bias

10-12. Data Science group projects

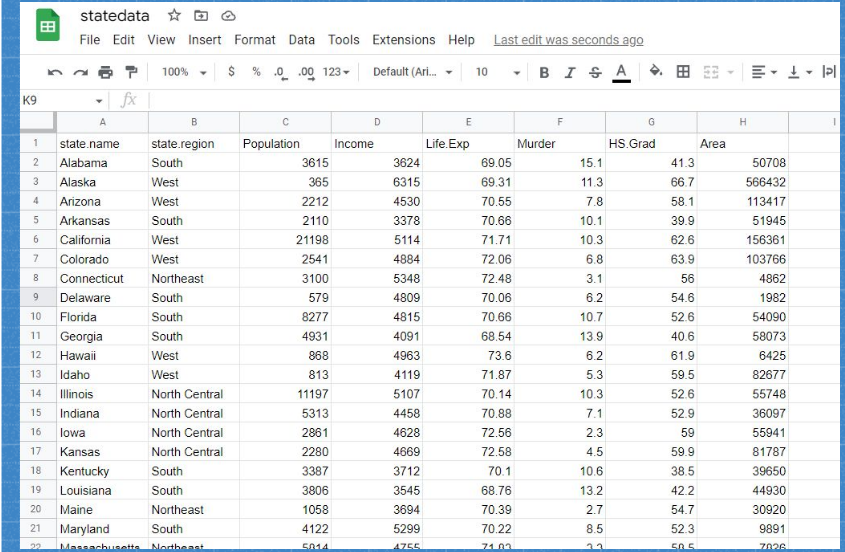
Standards: Data Analysis and Visualization

9-12.CT.2 Collect and evaluate data from multiple sources for use in a computational artifact.

9-12.CT.3 Refine and visualize complex data sets to tell different stories with the same data set.

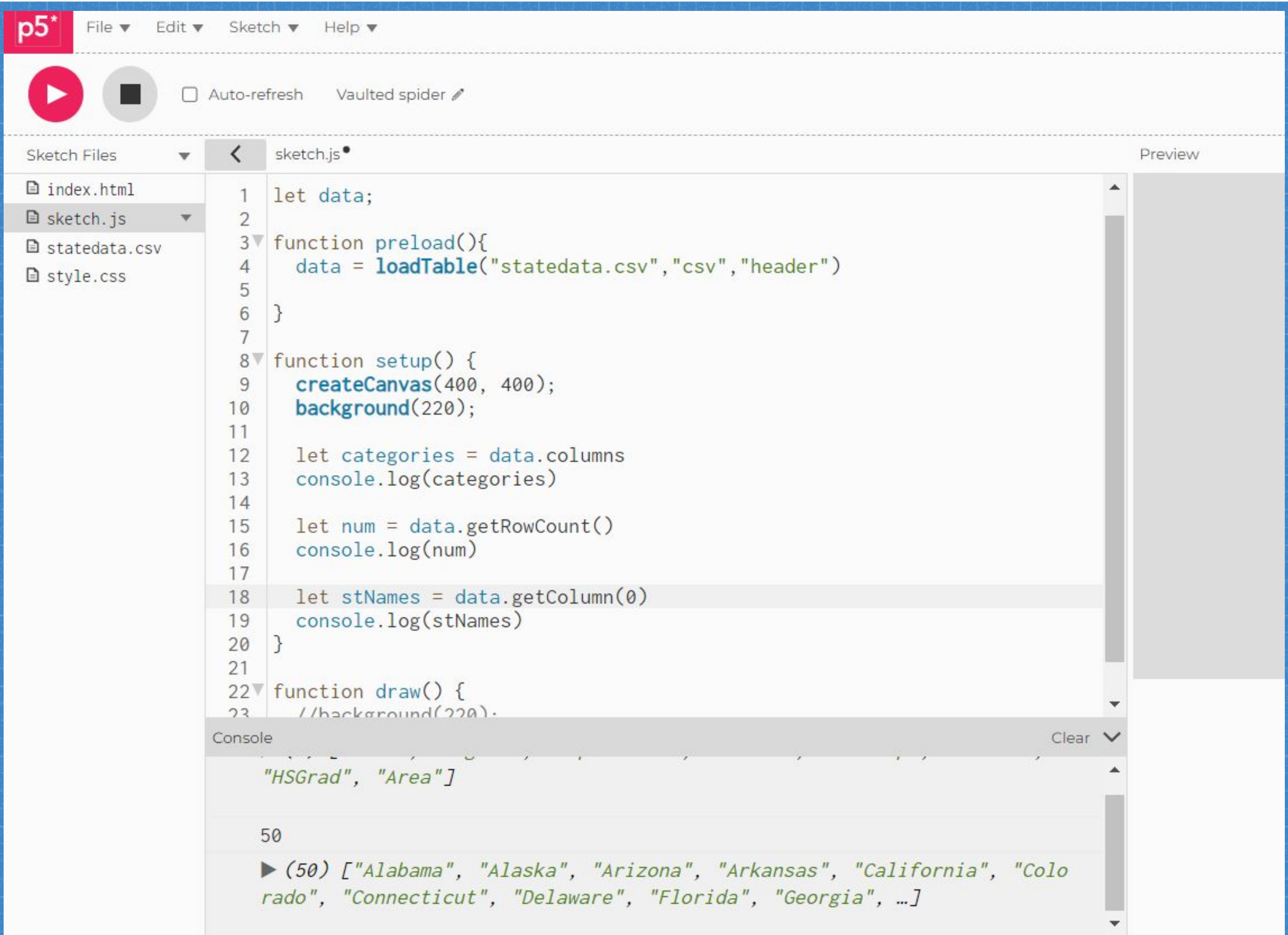
9-12.CT.7 Design or remix a program that utilizes a data structure to maintain changes to related pieces of data.

1) Intro to Data Science + importing data into p5.



	A	B	C	D	E	F	G	H
1	state.name	state.region	Population	Income	Life.Exp	Murder	HS.Grad	Area
2	Alabama	South	3615	3624	69.05	15.1	41.3	50708
3	Alaska	West	365	6315	69.31	11.3	66.7	566432
4	Arizona	West	2212	4530	70.55	7.8	58.1	113417
5	Arkansas	South	2110	3378	70.66	10.1	39.9	51945
6	California	West	21198	5114	71.71	10.3	62.6	156361
7	Colorado	West	2541	4884	72.06	6.8	63.9	103766
8	Connecticut	Northeast	3100	5348	72.48	3.1	56	4862
9	Delaware	South	579	4809	70.06	6.2	54.6	1982
10	Florida	South	8277	4815	70.66	10.7	52.6	54090
11	Georgia	South	4931	4091	68.54	13.9	40.6	58073
12	Hawaii	West	868	4963	73.6	6.2	61.9	6425
13	Idaho	West	813	4119	71.87	5.3	59.5	82677
14	Illinois	North Central	11197	5107	70.14	10.3	52.6	55748
15	Indiana	North Central	5313	4458	70.88	7.1	52.9	36097
16	Iowa	North Central	2861	4628	72.56	2.3	59	55941
17	Kansas	North Central	2280	4669	72.58	4.5	59.9	81787
18	Kentucky	South	3387	3712	70.1	10.6	38.5	39650
19	Louisiana	South	3806	3545	68.76	13.2	42.2	44930
20	Maine	Northeast	1058	3694	70.39	2.7	54.7	30920
21	Maryland	South	4122	5299	70.22	8.5	52.3	9891
22	Massachusetts	Northeast	5814	4755	71.83	3.3	58.5	7826

- What is data science?
- File types (pdf, txt, json, xls, csv)
- Spreadsheet to .csv
- loading csv into p5.js
- preload()
- loadTable()
- data.columns
- data.getRow()



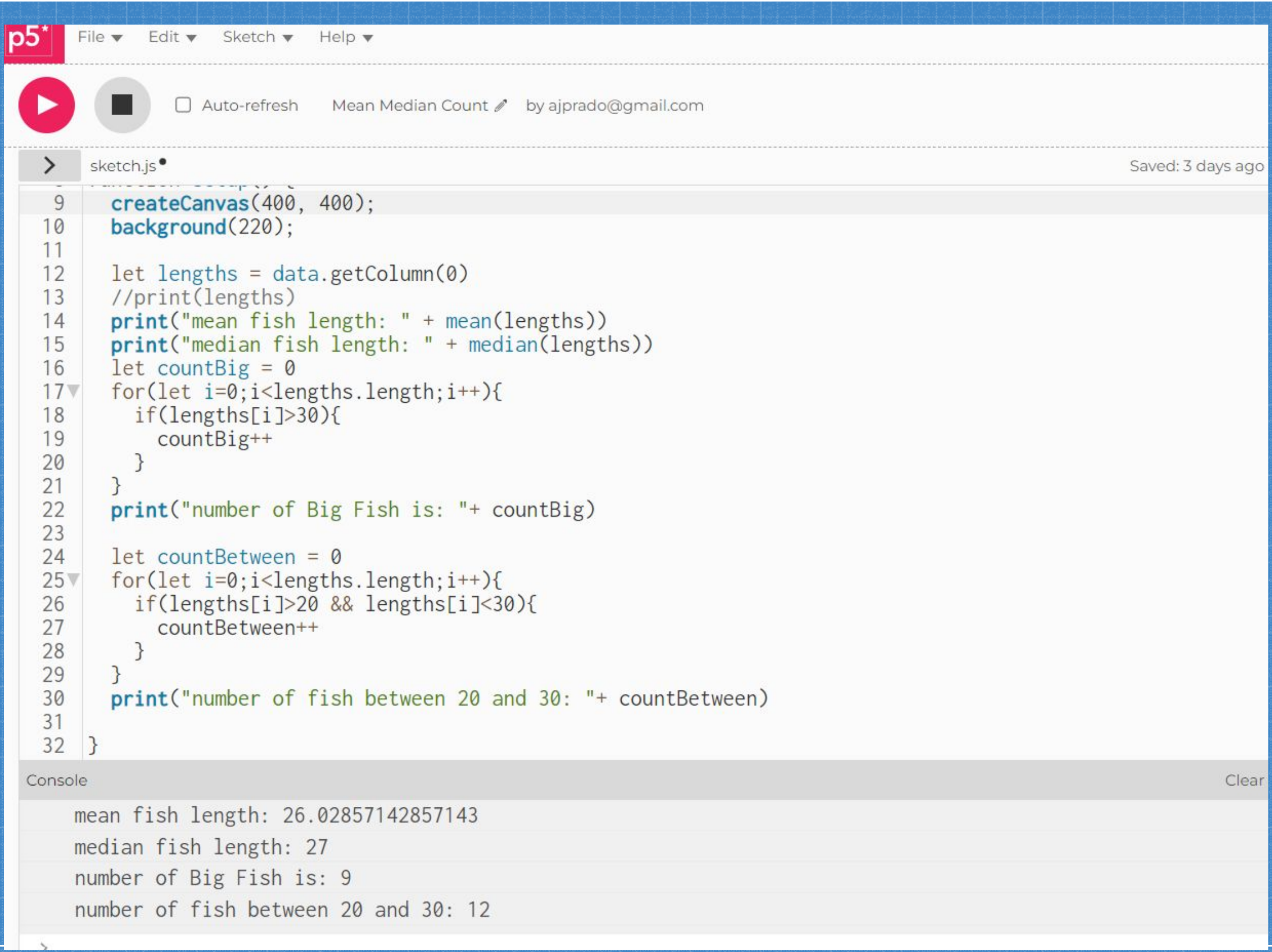
```
let data;

function preload(){
  data = loadTable("statedata.csv","csv","header")

}

function setup() {
  createCanvas(400, 400);
  background(220);

  let categories = data.columns
  console.log(categories)

  let num = data.getRowCount()
  console.log(num)

  let stNames = data.getColumn(0)
  console.log(stNames)
}

function draw() {
  //background(220);
```

```
"HSGrad", "Area"]

50
▶ (50) ["Alabama", "Alaska", "Arizona", "Arkansas", "California", "Colo
rado", "Connecticut", "Delaware", "Florida", "Georgia", …]
```

2) Creating functions with univariate data

- Discussion of Univariate Data
- Starter code with (Length of fish) and skeleton of comments with functions, instructions, and test conditions.
- Function to find mean
- Creating code to count values that meet a given condition
    - # fish over a certain size
    - # of fish between a range of 2 values

*) Function to find median, including .sort()



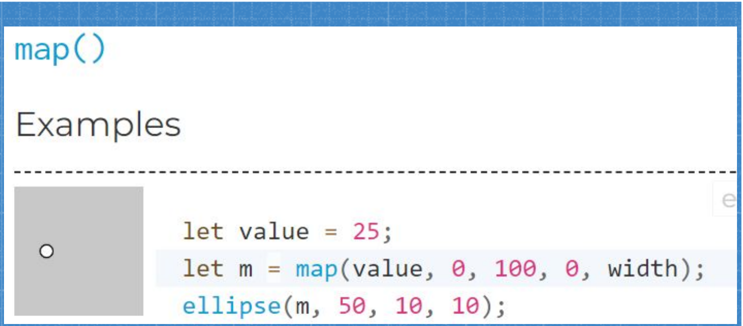
```
function setup() {
  createCanvas(400, 400);
  background(220);

  let lengths = data.getColumn(0)
  //print(lengths)
  print("mean fish length: " + mean(lengths))
  print("median fish length: " + median(lengths))
  let countBig = 0
  for(let i=0;i<lengths.length;i++){
    if(lengths[i]>30){
      countBig++
    }
  }
  print("number of Big Fish is: "+ countBig)

  let countBetween = 0
  for(let i=0;i<lengths.length;i++){
    if(lengths[i]>20 && lengths[i]<30){
      countBetween++
    }
  }
  print("number of fish between 20 and 30: "+ countBetween)

}
```

```
mean fish length: 26.02857142857143
median fish length: 27
number of Big Fish is: 9
number of fish between 20 and 30: 12
```

*. Cleaning and filtering data sets

- What are some of the problems with collecting data?
- What does cleaning data mean?
- Using a spreadsheet to look over data
- Changing heading names
- Removing entries with missing/erroneous values.
- Formatting data to work best in p5.js

3. Bar Graph (map ())



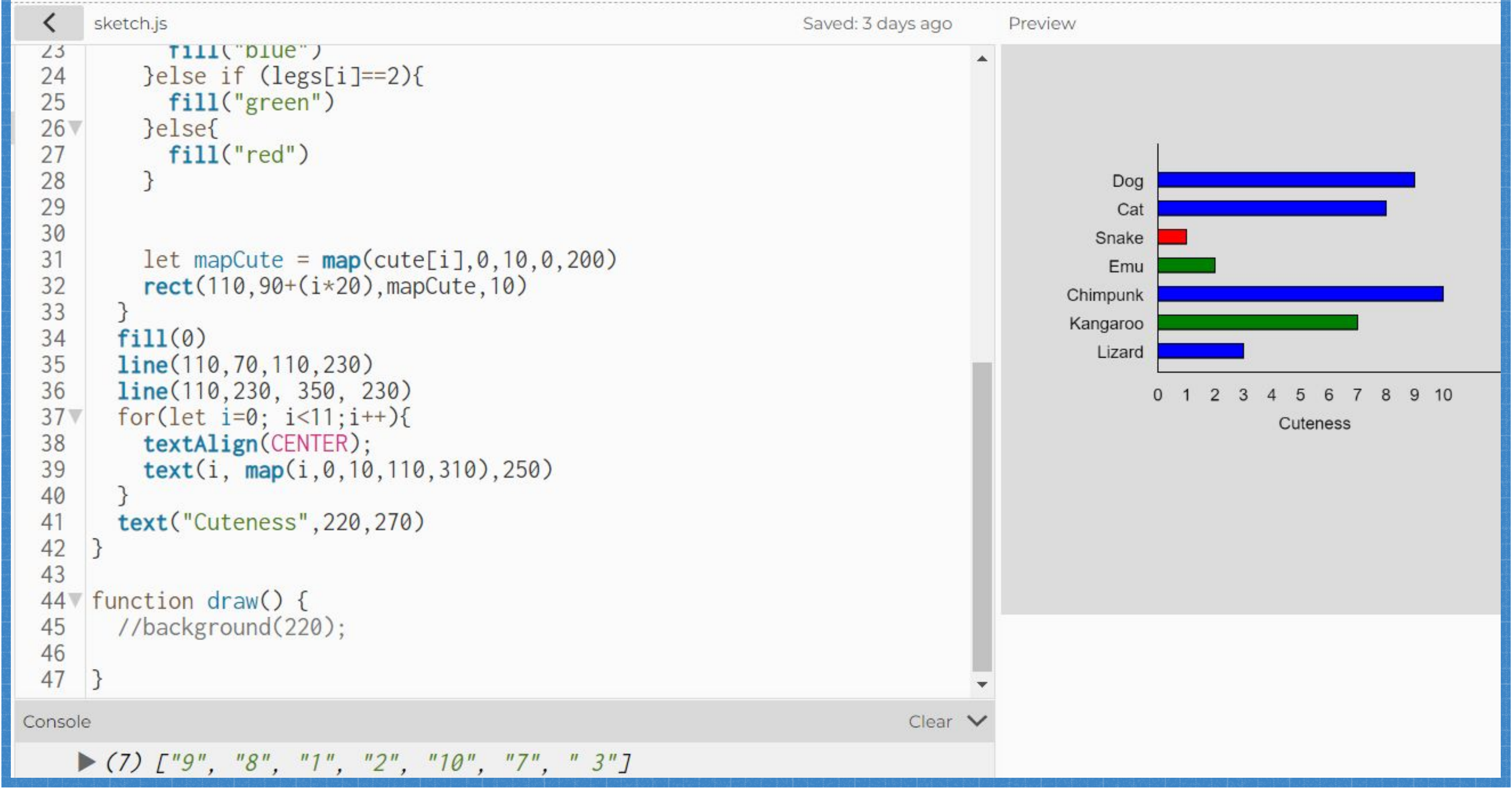
map()

Examples

```
let value = 25;
let m = map(value, 0, 100, 0, width);
ellipse(m, 50, 10, 10);
```

- Starter code with examples of using p5 map() function
- Practice using map() to edit the scale of numbers
- Starter code with data of animals and “cuteness” rating. (possibly scaffolded with axis set up)
- Use map to convert cuteness from 0 to 10 to a scale that will look better on the canvas.

*) Color the bars of the graph differently depending on the number of legs the animal has.

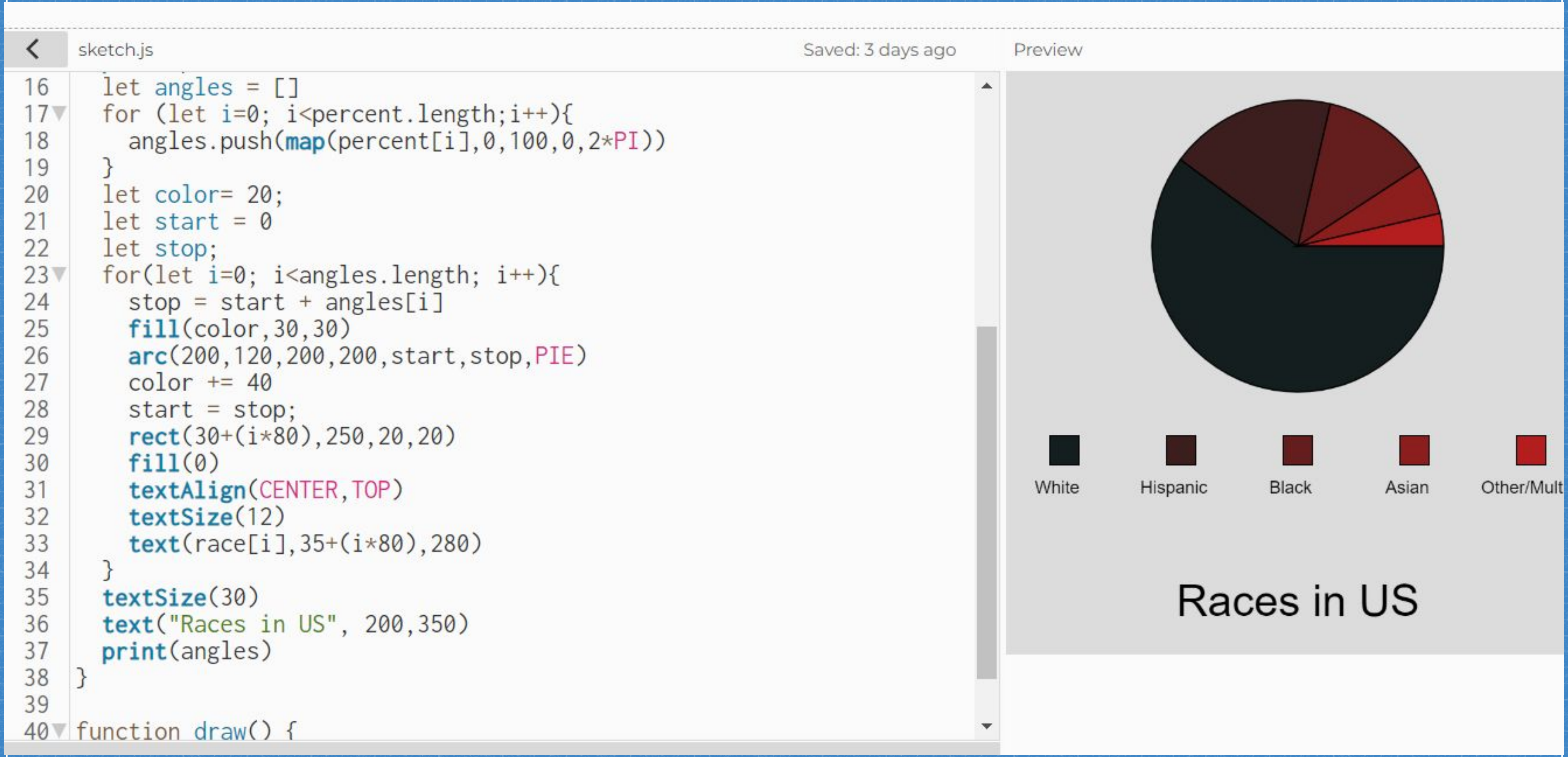


```
      fill("blue")
    }else if (legs[i]==2){
      fill("green")
    }else{
      fill("red")
    }


    let mapCute = map(cute[i],0,10,0,200)
    rect(110,90+(i*20),mapCute,10)
  }
  fill(0)
  line(110,70,110,230)
  line(110,230, 350, 230)
  for(let i=0; i<11;i++){
    textAlign(CENTER);
    text(i, map(i,0,10,110,310),250)
  }
  text("Cuteness",220,270)
}

function draw() {
  //background(220);

}
```

```
▶ (7) ["9", "8", "1", "2", "10", "7", " 3"]
```

4. Circle Graph

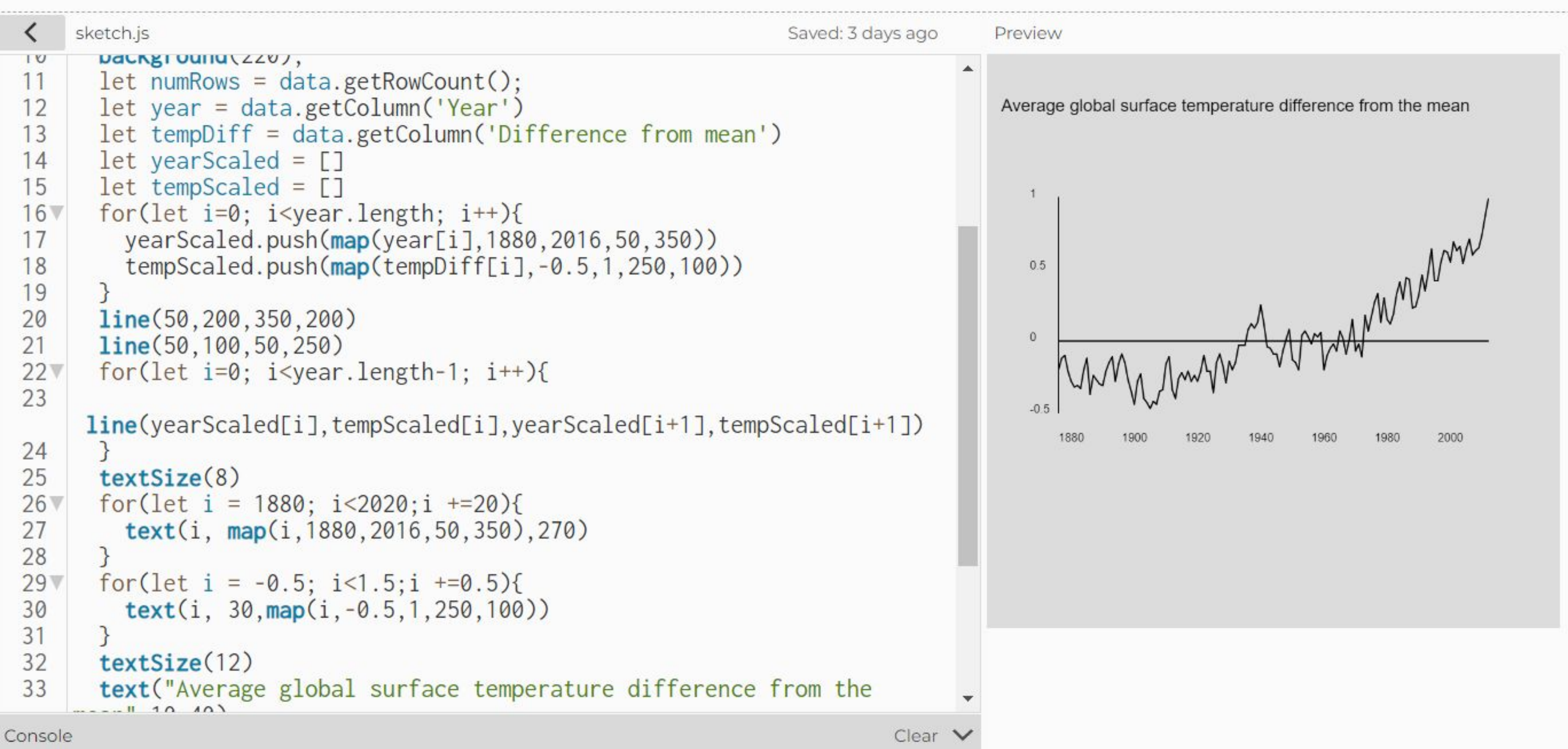
- Starter code with racial demographics of US.
- map the % for each race from 0 to 2PI to get the angle for each sector(brief explanation of radian)
- use arc() with angle measures to construct circle graph. (more technical so likely code-along)
- Create title and set up key to label each section



```
  let angles = []
  for (let i=0; i<percent.length;i++){
    angles.push(map(percent[i],0,100,0,2*PI))
  }
  let color= 20;
  let start = 0
  let stop;
  for(let i=0; i<angles.length; i++){
    stop = start + angles[i]
    fill(color,30,30)
    arc(200,120,200,200,start,stop,PIE)
    color += 40
    start = stop;
    rect(30+(i*80),250,20,20)
    fill(0)
    textAlign(CENTER,TOP)
    textSize(12)
    text(race[i],35+(i*80),280)
  }
  textSize(30)
  text("Races in US", 200,350)
  print(angles)
}

function draw() {
```

5. Bivariate Data + Line Graph

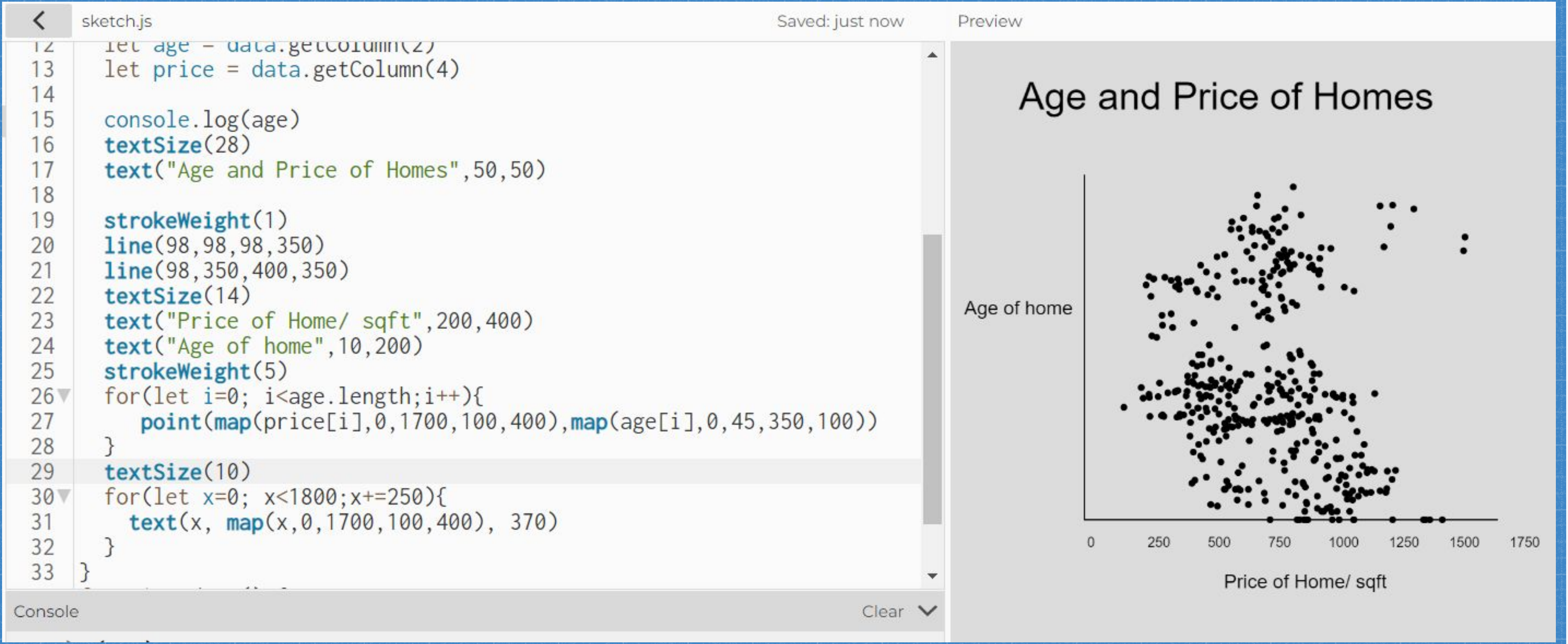
- What is bivariate data?
- How can we use data pairs to show change over time?
- Starter code with data for temperature difference from the mean for each month since 1880
- Code comments leading students to create the line graph.
- Use line() with map() to connect consecutive points (scaffolded hints for how to set up each map() scale conversion.



```
  background(220);
  let numRows = data.getRowCount();
  let year = data.getColumn('Year')
  let tempDiff = data.getColumn('Difference from mean')
  let yearScaled = []
  let tempScaled = []
  for(let i=0; i<year.length; i++){
    yearScaled.push(map(year[i],1880,2016,50,350))
    tempScaled.push(map(tempDiff[i],-0.5,1,250,100))
  }
  line(50,200,350,200)
  line(50,100,50,250)
  for(let i=0; i<year.length-1; i++){
    line(yearScaled[i],tempScaled[i],yearScaled[i+1],tempScaled[i+1])
  }
  textSize(8)
  for(let i = 1880; i<2020;i +=20){
    text(i, map(i,1880,2016,50,350),270)
  }
  for(let i = -0.5; i<1.5;i +=0.5){
    text(i, 30,map(i,-0.5,1,250,100))
  }
  textSize(12)
  text("Average global surface temperature difference from the mean", 10, 40)
```

6. Scatter Plot

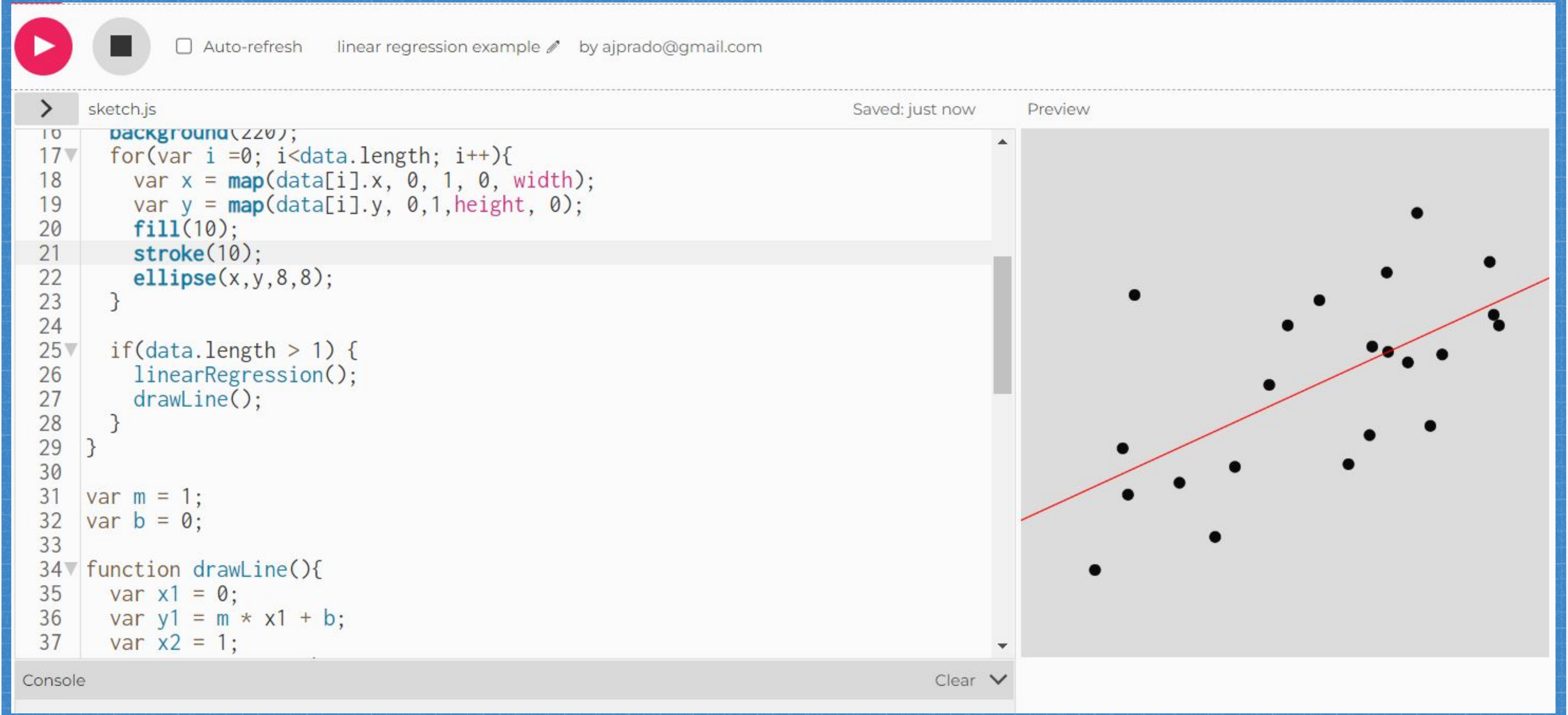
- Starter code comparing home prices and age of home.
- Code skeleton with comments leading students through the activity. Similar to line graph but using point() instead of lines.
- Discussion on what information can be found from the scatterplot.



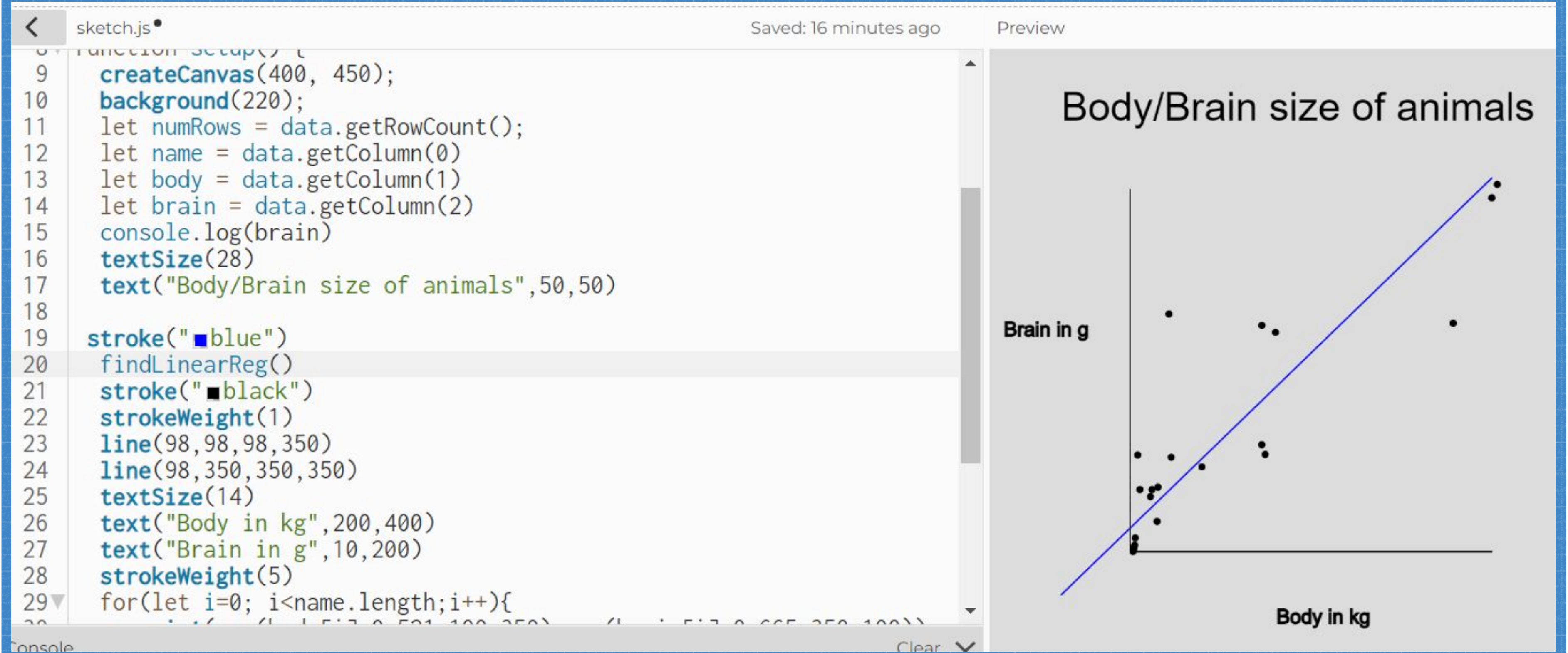
```
  let age = data.getColumn(2)
  let price = data.getColumn(4)

  console.log(age)
  textSize(28)
  text("Age and Price of Homes",50,50)

  strokeWeight(1)
  line(98,98,98,350)
  line(98,350,400,350)
  textSize(14)
  text("Price of Home/ sqft",200,400)
  text("Age of home",10,200)
  strokeWeight(5)
  for(let i=0; i<age.length;i++){
     point(map(price[i],0,1700,100,400),map(age[i],0,45,350,100))
  }
  textSize(10)
  for(let x=0; x<1800;x+=250){
    text(x, map(x,0,1700,100,400), 370)
  }
}
```

7. Linear regression

- Starting with a fully running program that allows students to plot points and it creates the trend line.
- Discussion of what is the meaning of the line.
- Strong vs. weak correlation examples
- Negative vs. positive correlation examples
- Given a mostly working data set with a prebuilt regression line calculator and having students finish it up with code comments.



```
  background(220);
  for(var i =0; i<data.length; i++){
    var x = map(data[i].x, 0, 1, 0, width);
    var y = map(data[i].y, 0,1,height, 0);
    fill(10);
    stroke(10);
    ellipse(x,y,8,8);
  }

  if(data.length > 1) {
    linearRegression();
    drawLine();
  }
}

var m = 1;
var b = 0;

function drawLine(){
  var x1 = 0;
  var y1 = m * x1 + b;
  var x2 = 1;
```

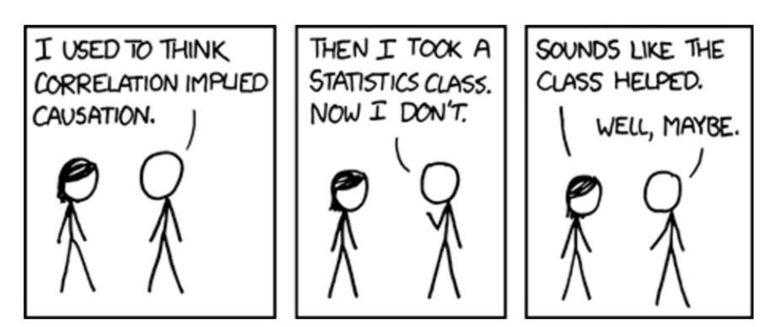


```
function setup() {
  createCanvas(400, 450);
  background(220);
  let numRows = data.getRowCount();
  let name = data.getColumn(0)
  let body = data.getColumn(1)
  let brain = data.getColumn(2)
  console.log(brain)
  textSize(28)
  text("Body/Brain size of animals",50,50)

 stroke("blue")
  findLinearReg()
  stroke("black")
  strokeWeight(1)
  line(98,98,98,350)
  line(98,350,350,350)
  textSize(14)
  text("Body in kg",200,400)
  text("Brain in g",10,200)
  strokeWeight(5)
  for(let i=0; i<name.length;i++){
```

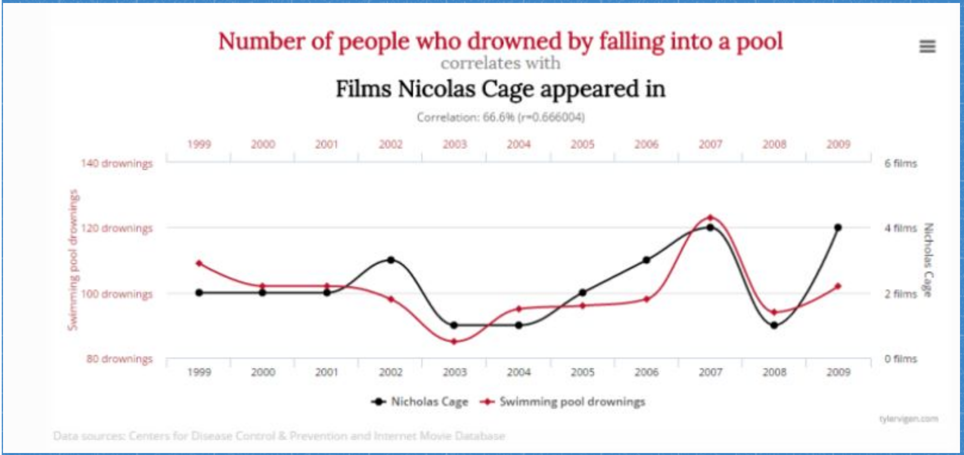
8. Analyzing results and making conclusions

- Activity with examples of ridiculous correlated data. Discussion of the important difference between correlation and causation.

- Example of data conclusions using overgeneralization, discussion the scope of conclusions that can be made.



I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.



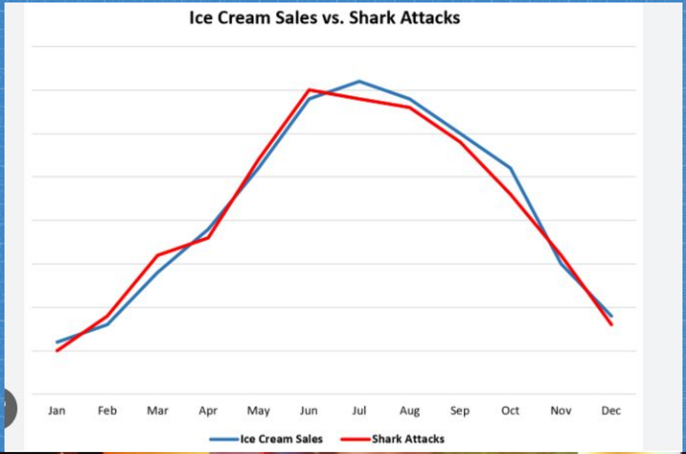
Number of people who drowned by falling into a pool

correlates with

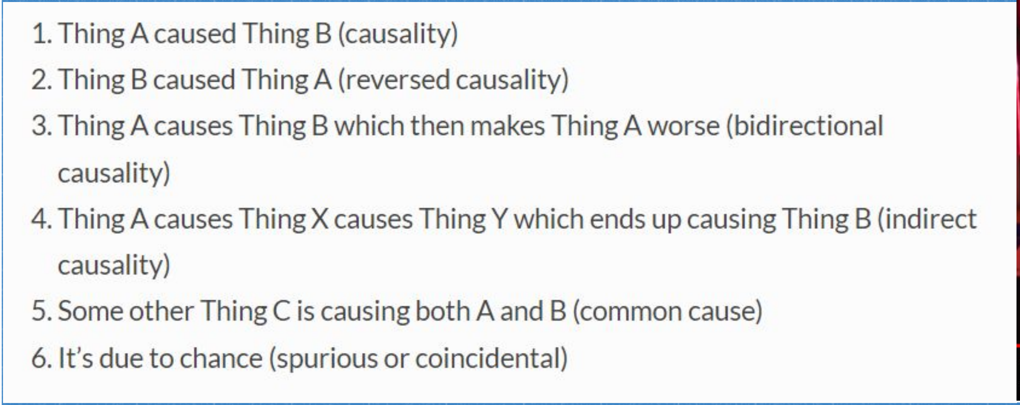
Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database



Ice Cream Sales vs. Shark Attacks

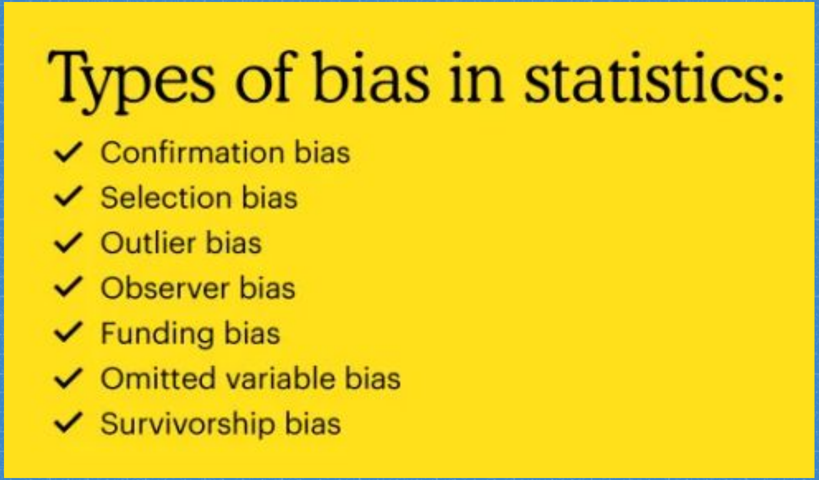


1. Thing A caused Thing B (causality)
2. Thing B caused Thing A (reversed causality)
3. Thing A causes Thing B which then makes Thing A worse (bidirectional causality)
4. Thing A causes Thing X causes Thing Y which ends up causing Thing B (indirect causality)
5. Some other Thing C is causing both A and B (common cause)
6. It's due to chance (spurious or coincidental)

In Nature there was a study in 1999 that showed

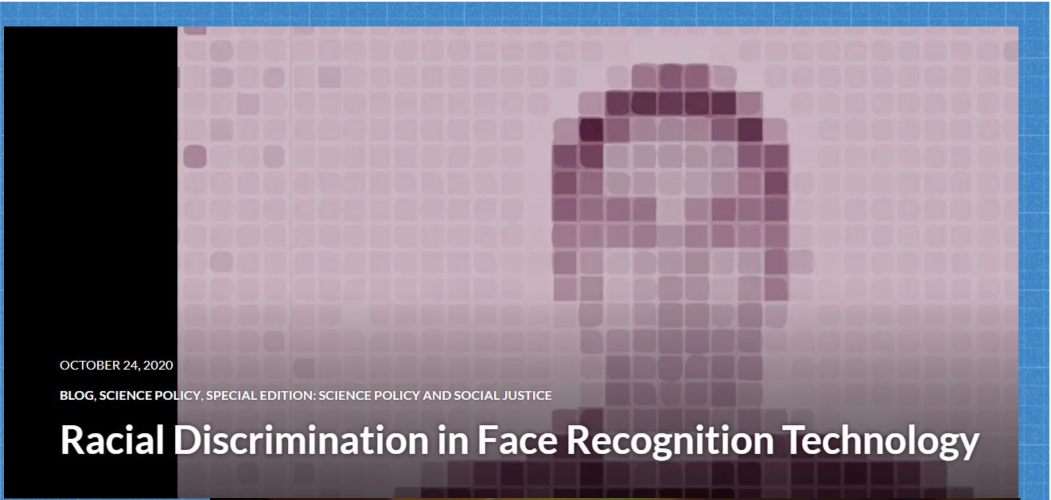
9. Concerns about bias

- Video regarding concerns of bias data/conclusions.
- Article with examples of several of the types of statistical bias.
- Discussion of reliability of data (random sample, collection techniques, inclusive data …)



Types of bias in statistics:

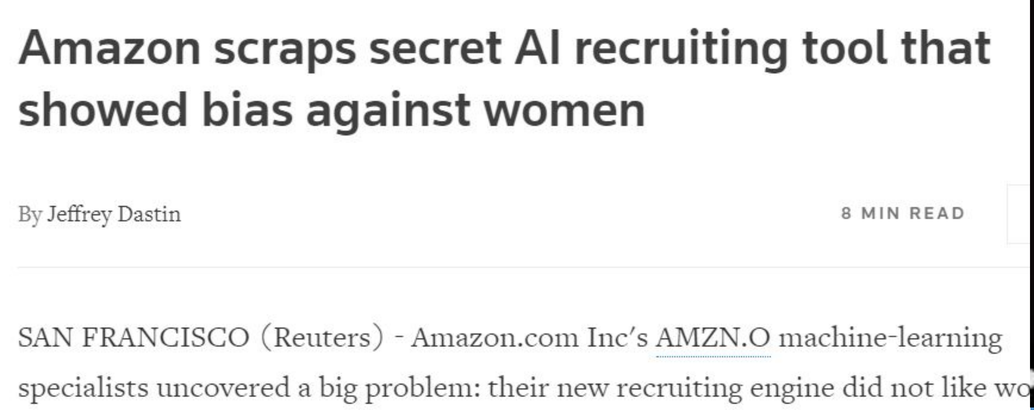
- ✓ Confirmation bias
- ✓ Selection bias
- ✓ Outlier bias
- ✓ Observer bias
- ✓ Funding bias
- ✓ Omitted variable bias
- ✓ Survivorship bias



OCTOBER 24, 2020

BLOG, SCIENCE POLICY, SPECIAL EDITION: SCIENCE POLICY AND SOCIAL JUSTICE

Racial Discrimination in Face Recognition Technology



Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like wo

10-12. Data Science group projects

- Find a data set you are interested in exploring, preferably something that impacts you or your community.

- Create at least two data visualizations from it using p5.

- Explain any conclusions you can make about the data including justification.

- Discuss any possible concerns about bias or errors?

- 5 minutes presentation to the class about what you discovered.