Introduction to Data Science - Final Exam

Directions:

Be sure to justify all of your answers.

Please ask if you do not understand something.

Short answer questions are just that -- **short answer**. Do not spend too much time on them.

If you get stuck, move on!

In [1]:
```
from datascience import *
import numpy as np

import matplotlib.pyplot as plots
plots.style.use('fivethirtyeight')
%matplotlib inline
```

In [2]:
```
def standard_units(any_numbers):
    "Convert any array of numbers to standard units."
    return (any_numbers - np.mean(any_numbers)) / np.std(any_numbers)

def correlation(t, x, y):
    """Return the correlation coefficient (r) of two variables."""
    return np.mean(standard_units(t.column(x)) * standard_units(t.column(y))

def slope(t, x, y):
    """The slope of the regression line (original units)."""
    r = correlation(t, x, y)
    return r * np.std(t.column(y)) / np.std(t.column(x))

def intercept(t, x, y):
    """The intercept of the regression line (original units)."""
    return np.mean(t.column(y)) - slope(t, x, y) * np.mean(t.column(x))

def fit(table, x, y):
    """Return the height of the regression line at each x value."""
    a = slope(table, x, y)
    b = intercept(table, x, y)
    return a * table.column(x) + b
```

The website you may use as a coding reference is HERE

Task 1

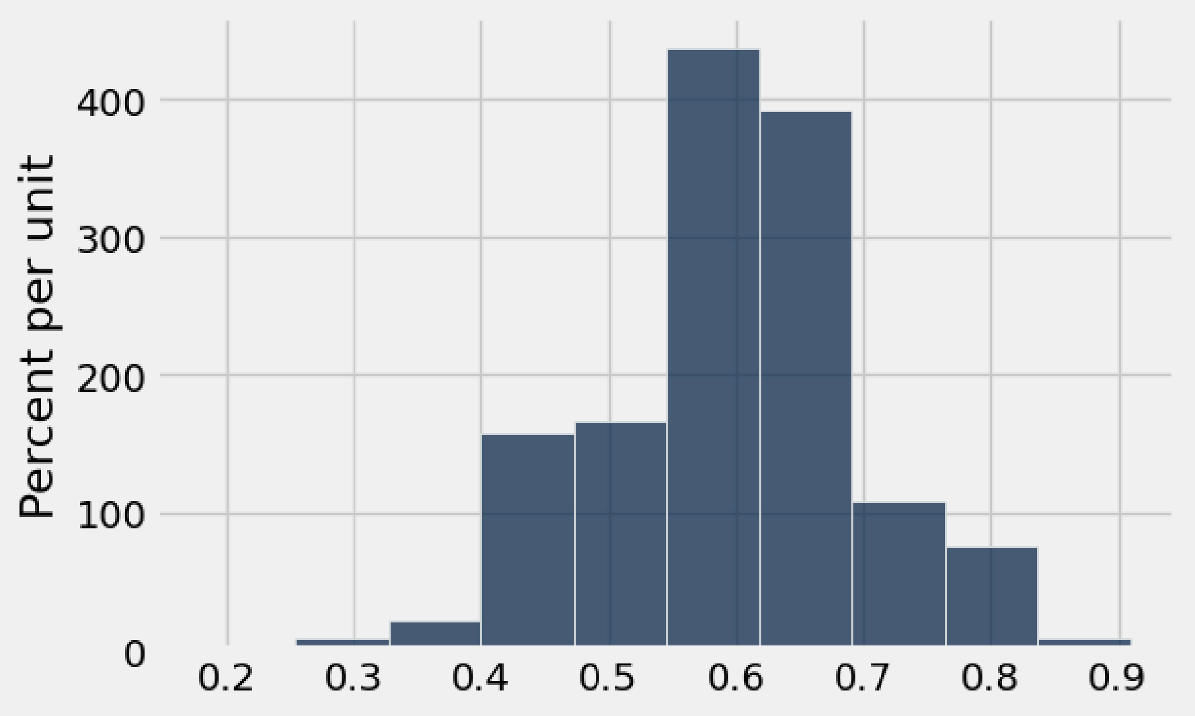
I have three pairs of pants that I choose from each morning--my gray farm pants, by blue jeans, and my Harry Potter pants. I tell my wife that I just randomly grab one of the three each morning. However, she thinks I secretly like my Harry Potter pants better. She takes data for 22 days and notes that I wear the Harry Potter pants 13 days.

Using this information, create a 95% confidence interval for the probability that I choose my Harry Potter pants each day.

In [6]:

```
pants = make_array(13/22, 9/22)
potter_prob = make_array()

for i in np.arange(5000):
    potter22 = sample_proportions(22, pants).item(0)
    potter_prob = np.append(potter_prob, potter22)

Table().with_column("", potter_prob).hist()

left = round(percentile(2.5, potter_prob)*100, 2)
right = round(percentile(97.5, potter_prob)*100, 2)
print ("The probability for the professor to choose his Harry Potter pants e
```

```
The probability for the professor to choose his Harry Potter pants each day
is in between 40.91 % and 77.27 % for a 95% confidence interval.
```



Short Answer 1

You can use your results in Task 1 to conduct a hypothesis test. What would the null hypothesis be and what could you say about the p-value? Can you make a decision based on your results? If so, what would it be?

`null hypothesis:` Kenneth choosing pants among three for each morning is just a random chance.

In [7]:

```
p_value = np.count_nonzero(potter_prob >= 13/22)/5000
p_value
```

Out[7]:

```
0.5884
```

Since the p-value is 0.5884, we can see that almost half of the sample show the Harry Potter pants have been chosen more than 13 days in 22 days, which is a lot. Therefore, we can say that the probability to choose the pants among three each morning is not just a random chance.

Task 2

I asked 100 randomly chosen college students who were not on althetic teams whether or not they drank alcohol more than once a month. I then asked them to see how many sit-ups they could do in 60 seconds. The data are in the table below.

In [8]:

```
situps = Table().read_table("~/DS_113_S23/Tests/situps.csv")
situps
```

Out[8]:

ID	Drink (Y/N)	Sit-ups
1	N	13
2	Y	23
3	Y	22
4	N	23
5	N	21
6	N	28
7	N	32
8	N	30
9	N	28
10	N	27

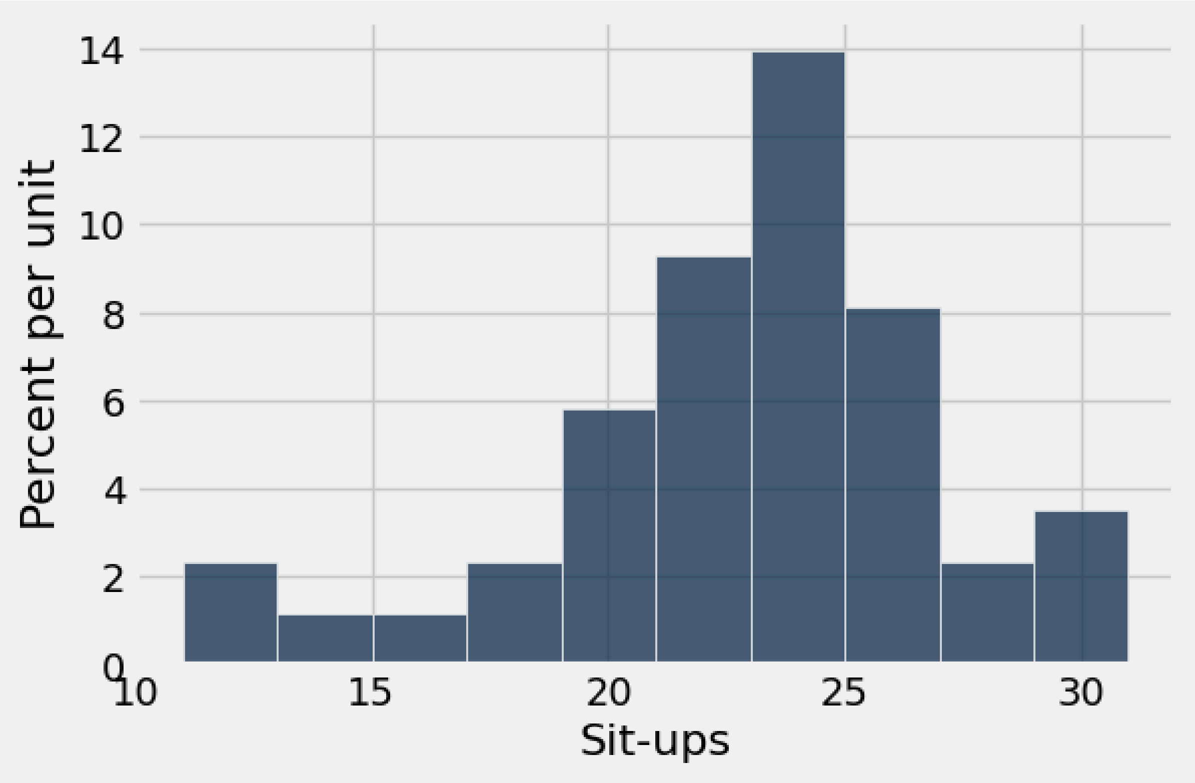
... (90 rows omitted)

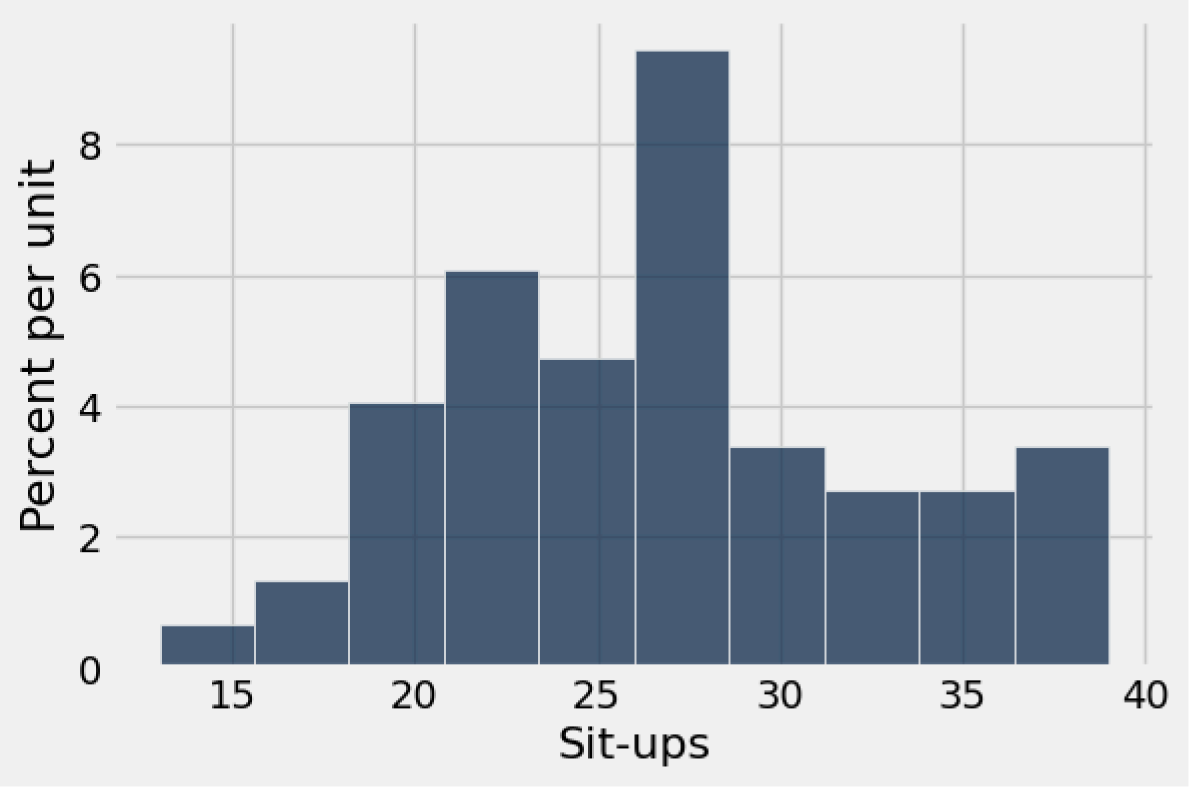
Task 2.A.

Create histograms for the number of sit-ups done by each group.

In [14]:

```
situps.drop(0).where("Drink (Y/N)", are.equal_to("Y")).hist("Sit-ups")
situps.drop(0).where("Drink (Y/N)", are.equal_to("N")).hist("Sit-ups")
```





Short Answer 2

Please estimate the standard deviation for the number of sit-ups by each group by looking at the histograms. **As part of your answer, please explain clearly what the standard deviation means and how you can estimate it from the histogram.**

The standard deviation is the square root of the variance and this shows how far each values are from the mean. Since the histogram of the people who do not drink is more spread than the histogram of the people who drink alcohol, we can know that the standard deviation, which is the distance from the mean and the actual value, is bigger.

The standard deviation of the people who drink would be about 5 and the people who do not drink would be about 8.

Task 2.B.

Please create a table with two columns and two rows. Column labels should be "Drink (Y/N)" and "Average number of sit-ups" and should show the appropriate information.

In [19]:

```
situp = situps.drop(0).group("Drink (Y/N)", np.average).relabel("Sit-ups ave
situp
```

Out[19]:

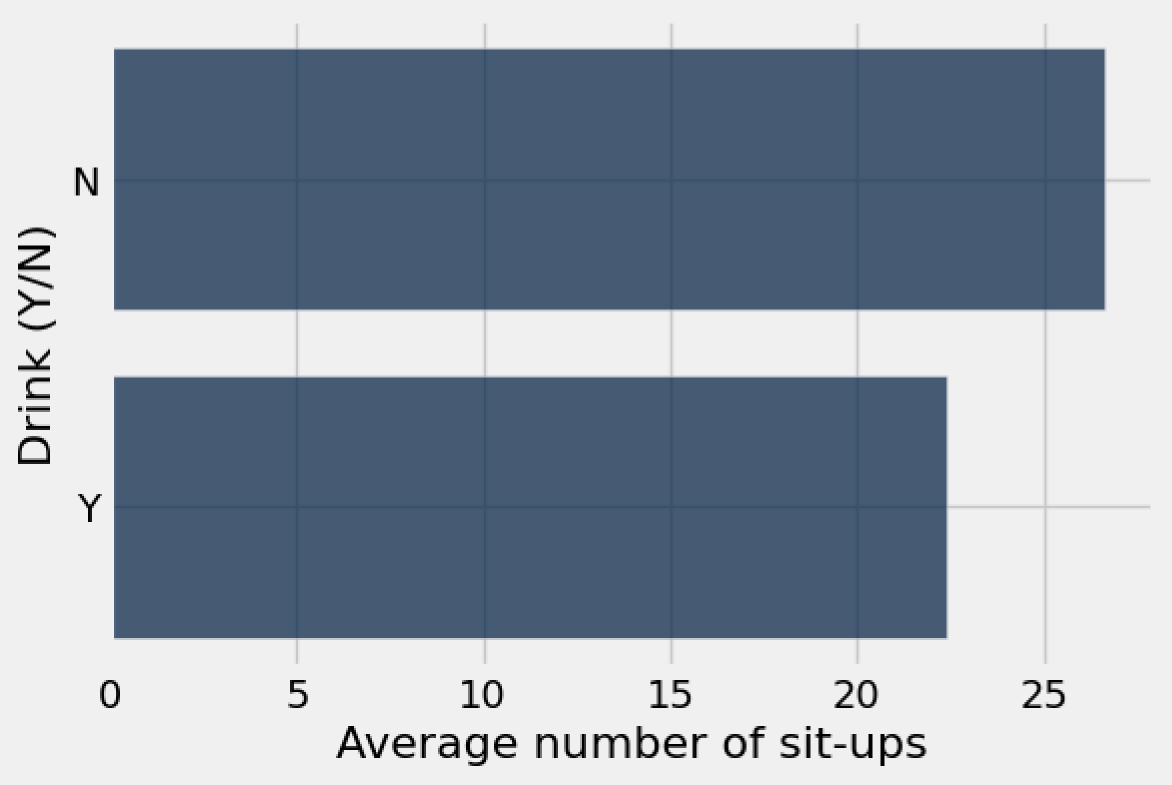
Drink (Y/N)	Average number of sit-ups
N	26.6316
Y	22.3721

Task 2.C.

Please make a bar chart showing the information in your table from part 2.B.

In [21]:

```
situp.barh("Drink (Y/N)")
```



Task 3

Using the data from Task 2, please conduct a simulation-based hypothesis test to determine if there is a difference between the number of sit-ups amongst students who drink and amongst students who do not drink (at least more than once a month).

As part of your analysis, please be sure to report the following:

A) Your null hypothesis

B) Your test statistic

C) Your p-value

D) Your conclusion

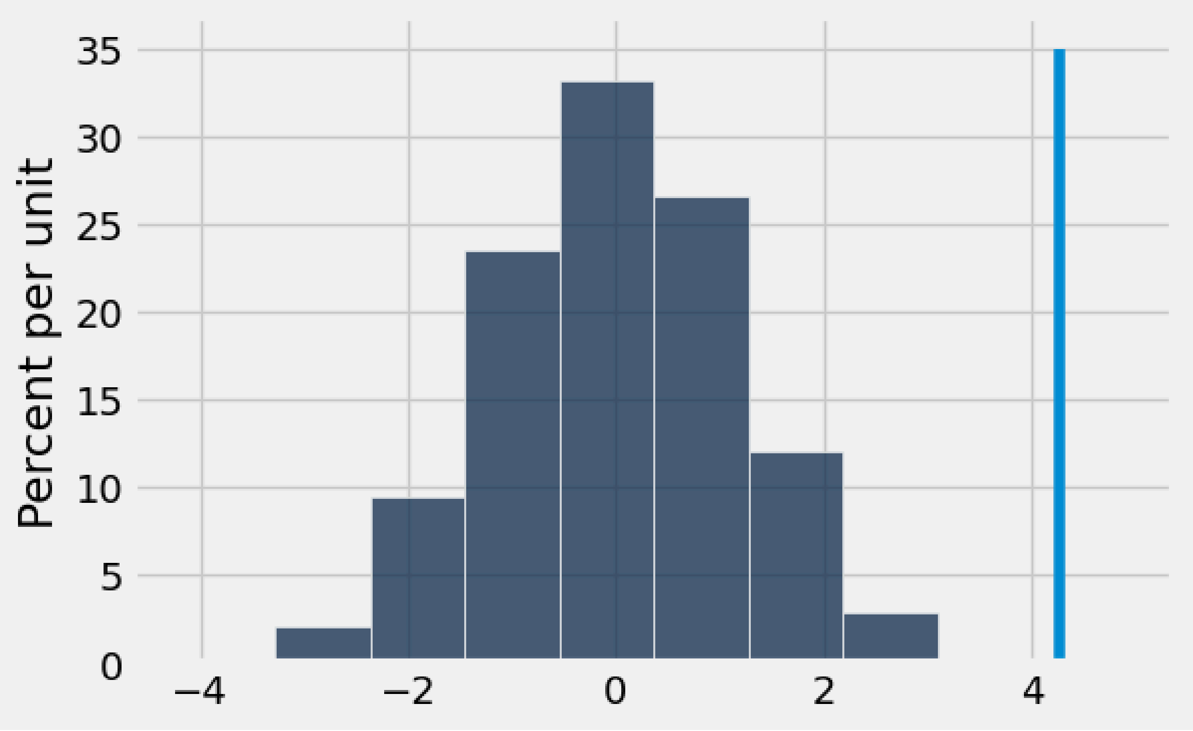
E) What your conclusion tells us about the two populations

`null hypothesis:` There is no factor that affects the number of sit-ups amongst students who drink and amongst students who do not drink.

`test statistic:` Difference in the average number of sit-ups of students who drink and who do not drink

In [50]:
```
observed = situp[1].item(0) - situp[1].item(1)

situp_array = situps[2]
drink_num = situps.where("Drink (Y/N)", "Y").num_rows
no_drink_num = situps.where("Drink (Y/N)", "N").num_rows
test_stat = make_array()

for i in np.arange(5000):
    shuffled = np.random.choice(situp_array, len(situp_array), replace=False
    drink_array = shuffled[:drink_num]
    drink_avg = np.average(drink_array)
    no_drink_array = shuffled[drink_num:]
    no_drink_avg = np.average(no_drink_array)
    test_stat = np.append(test_stat, no_drink_avg-drink_avg)

Table().with_column("", test_stat).hist()
plots.plot([observed,observed], [0,0.35])

p_value = np.count_nonzero(test_stat >= observed)/5000
print("p-value:", p_value)
```

```
p-value: 0.0002
```



When we shuffled the samples under the null hypothesis, since the p-value is smaller than 0.05, there is an evidence to reject the null hypothesis, which is that there is not factor that affects the number of sit-ups amongst students who drink and amongst students who do not drink. There is something other than just the random chance in the difference of the number of sit-ups in two groups.

Short Answer 3

What exactly does a p-value mean? If I use a p-value cut-off of 5% and the null hypothesis is true, what is the chance I make an incorrect decision?

Be sure to justify your answer.

A p-value is the probability to obtain the test statistics under the null hypothesis.

The chance to make an incorrect decision when the null is true is 5% when the p-value cut-off is 5%.

Task 4

The data table below shows the nasal length and nasal width in millimeters for a sample of kangaroos.

In [51]:

```
kangaroos = Table().read_table("~/DS_113_S23/Tests/Kangaroos.csv")
kangaroos
```

Out[51]:

ID	Nasal Length (mm)	Nasal Width (mm)
1	609	241
2	629	222
3	620	233
4	564	207
5	645	247
6	493	189
7	606	226
8	660	240
9	630	215
10	672	231

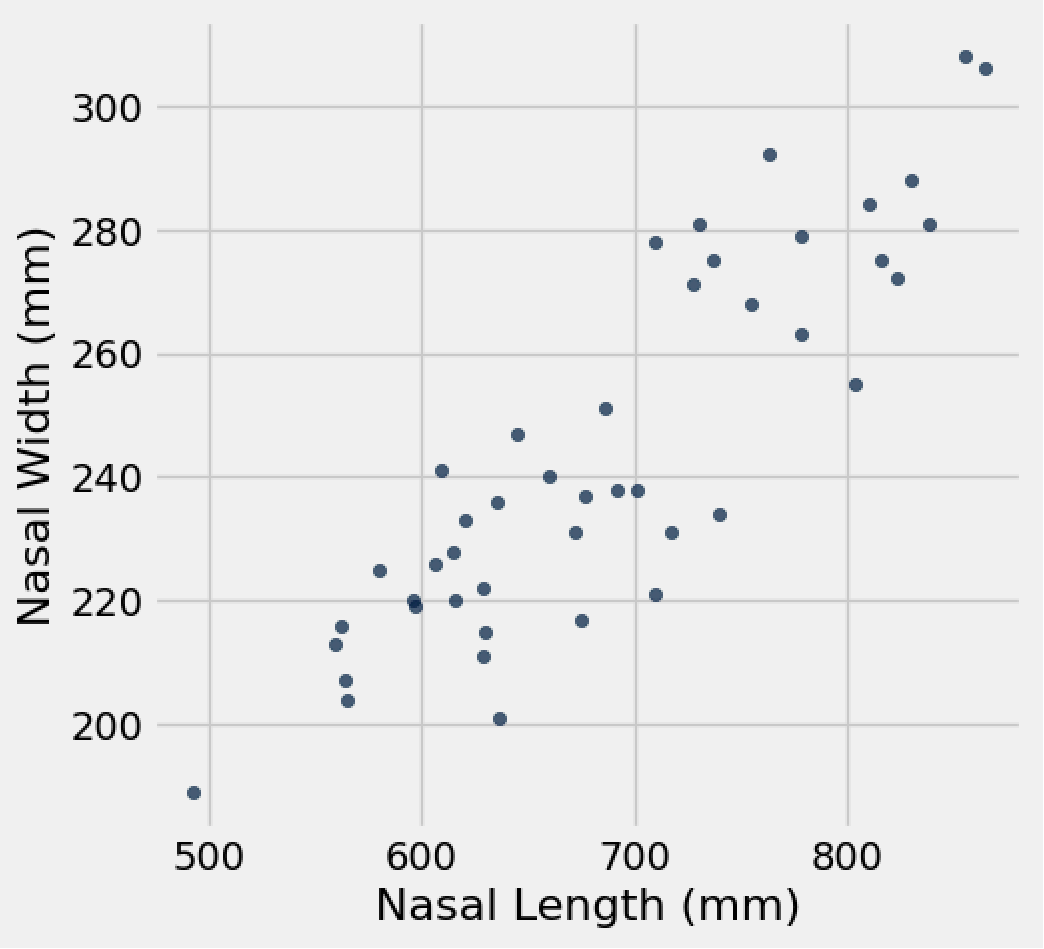
... (35 rows omitted)

Task 4.A.

Please create a scatter plot of the data and interpret what you see.

In [52]:

```
kangaroos.scatter("Nasal Length (mm)", "Nasal Width (mm)")
```



There is a positive correlation between the nasal length and the nasal width. Usually, kangaroos with the longer nasal length have the wider nasal width.

Short Answer 4

Estimate the correlation coefficient for these data and justify your answer. How would the data change if the correlation coefficient were higher? How would it change if the sign changed (i.e. you multiplied it by -1)?

`estimated correlation codfficient:` 0.8

- the data are scattered along a linear line and they are kind of crowded

If the correlation coefficient were higher, than the kangaroos with similar nasal length would have almost the same nasal width. And if the sign changed for the correlation coeffienct, kangaroos with the longer nasal length will have the narrower nasal width.

Task 4.B.

Find the regression line relating nasal width and length and use it to predict the nasal width of a kangaroo with a nasal length of 616 mm.

In [56]:
```
sl = slope(kangaroos, "Nasal Length (mm)", "Nasal Width (mm)")
it = intercept(kangaroos, "Nasal Length (mm)", "Nasal Width (mm)")

prediction = sl*616 + it
prediction
```

```
Out[56]: 223.62003246311374
```

A kangaroo with a nasal length of 616mm is predicted to have the nasal width about 223.62mm.

Task 4.C.

Add a column to the table "kangaroos" that contains the predicted nasal widths for each kangaroo according to our linear model.

```
In [57]: kangaroos = kangaroos.with_column("Predicted Nasal Width", fit(kangaroos, "N
         kangaroos
```

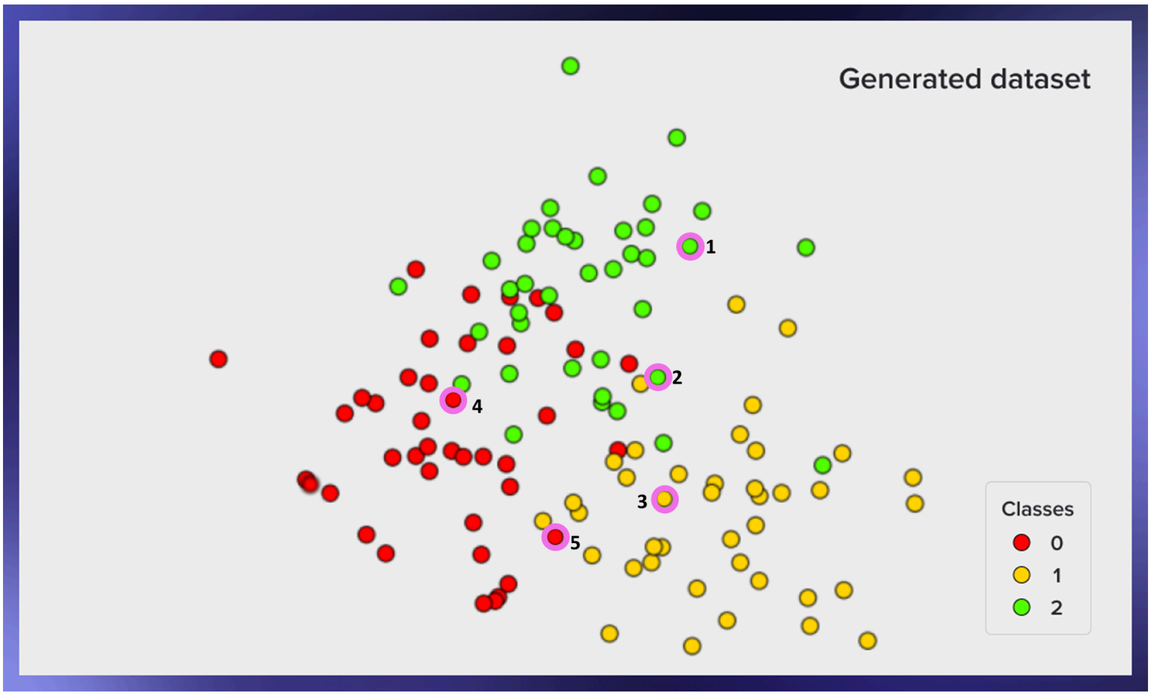
Out[57]:

ID	Nasal Length (mm)	Nasal Width (mm)	Predicted Nasal Width
1	609	241	221.607
2	629	222	227.359
3	620	233	224.77
4	564	207	208.664
5	645	247	231.961
6	493	189	188.244
7	606	226	220.744
8	660	240	236.275
9	630	215	227.647
10	672	231	239.726

... (35 rows omitted)

Short Answer 5

This figure displays samples that have had three variables measured about them. Two are quantitative and shown in standardized units on the x- and y-axis. The third is categorical and comprised of three categories.



Please categorize the five labeled test samples using the nearest neighbor classification algorithm and the 5-nearest-neighbor classification algorithms. Which algorithm is more accurate?

Nearest Neighbor:

1 = 2

2 = 1

3 = 1

4 = 2

5 = 1

5 - Nearest Neighbor:

1 = 2

2 = 2

3 = 1

4 = 0

5 = 1

The 5-nearest-neighbor classification algorithms is more accurate.

Short Answer 6

Based on the data in question 5, which of the following values do you think would be closest to the accuracy of the nearest neighbor classification algorithm: 20%, 50%, or 90%?

To receive credit, you must justify your answer.

If we use the 50% or 90% nearest neighbor classification algorithm, then the accuracy would not be precise. If we choose to use 90% in the algorithm, then almost all of the values in the data would be included and this would cuase most of the values to be labeled as the majority class. 50% is also not recommended, since the values in the data are crowded in the middle and if we choose 50% to use in the algorithm, it would cause the majority class labeled for the most of the values in the middle. Therefore, I think 20% would be closest to the accuracy of the nearest neighbor classification algorithm.

Extra Challenge Task

Suppose a lion chooses between two different prey each night. She either hunts gazelle or flamingo. When she hunts flamingo, she is successful 45% of the time. When she hunts gazelle, she is successful 25% of the time. Your job is to determine if she has a preferred prey--that is one that she hunts more often than the other.

You observe her returning from the hunt for 50 days. She brings back 8 gazelle and 9 flamingoes and nothing the other nights. Please describe how you would do a simulation test to determine if she has a preferred prey. If you have time, go ahead and conduct the test.

In [59]:

```
gazelle_success = make_array(0.25, 0.75)
flamingo_success = make_array(0.45, 0.55)
day50_success = make_array(8/50, 9/50, 33/50)
```

In [ ]: