Welcome to

Learn to Code

wifi: galvanize guest seattle
(no password, social sign in)
(Pro-Tip: go to bing.com instead of google.com)
Current Version: March 23rd 2017

powered by galvanize

About Galvanize

Dynamic learning community for technology

- Web Development
- Workspace
- Data Science
- Networking

To learn more, visit galvanize.com

powered by galvanize

Learn Data Science with Galvanize

Data Science Fundamentals: Intro to Python

- 6 week part-time workshop

Data Science Immersive Program

- 12 week full-time program

GalvanizeU

- 12 month program in San Francisco
- Fully-accredited by the University of New Haven

To learn more, visit galvanize.com/data-science
Or email enrollment@galvanize.com

powered by galvanize

For more information

Email **Lee Ngo** at
lee.ngo@galvanize.com
or
Visit our website at
galvanize.com

powered by galvanize

But first...

powered by galvanize

Let’s get to know each other

Turn to the person next to you and ask:

1) What is your name?
2) Why did you come here?
3) What is one mystery you’d like to investigate if you could?

You have 2 minutes to complete this mission!

Intro to Data Science

Using Python

powered by galvanize

About this Workshop's Architect

Matt Drury

github.com/madrury

Lead Instructor & Principal Data Scientist @ Galvanize

Usually uses Spaceman Spiff as an avatar

About this Workshop's Instructor

Lee Ngo

github/lee-ngo

Galvanize Evangelist based in Seattle

Once did a Poisson regression on geolocation data

About this Workshop's Instructor

Mari Pierce-Quinonez

github.com/maripqz

gStudent - Data Science

Trying out new recipes and talking to her houseplants

About this Workshop's Instructor

Brian McAdams

github.com/theastrocatgStudent - Data Science

"I do data things. I have a beard. Portland."

In this course you will learn

- ❑ Set up your computer for Jupyter Notebook
- ❑ Importing Libraries
- ❑ Loading and Inspecting Data
- ❑ Creating Visualizations
- ❑ Creating a Linear Regression

Pre-requisite courses

- ❑ Intro to Python for Data Science
- ❑ Explorations in Python for Data Science

OK if you have zero exposure, but recommended to return when these courses launch again

Gut check, Galvanize style!

- This course is for beginners
- Feel free to move ahead
- Help others when you can
- Be patient and nice
- We'll all get through it!

Want to move ahead? No problem!

Go to: github.com/madrury-Galvanize/learn-to-code-data-science/

Or: bit.ly/madrury-ltc-ds

Clone, fork or download the repo!

Setting up your computer

(Brace yourself…)

1: Install Anaconda!

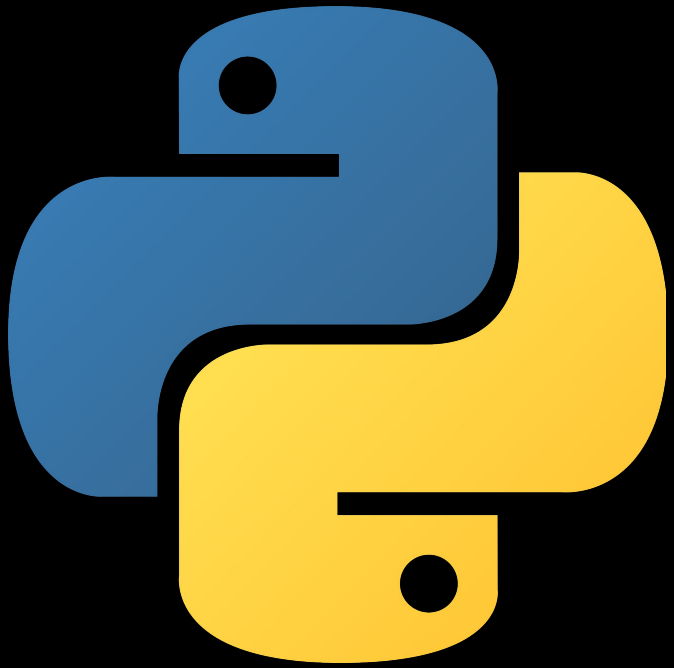
continuum.io/downloads

Download here ^

Follow the instructions in the website - they vary per platform

Anaconda is an open-source platform for Python, powered by Continuum Analytics.

Anaconda Installs Python!



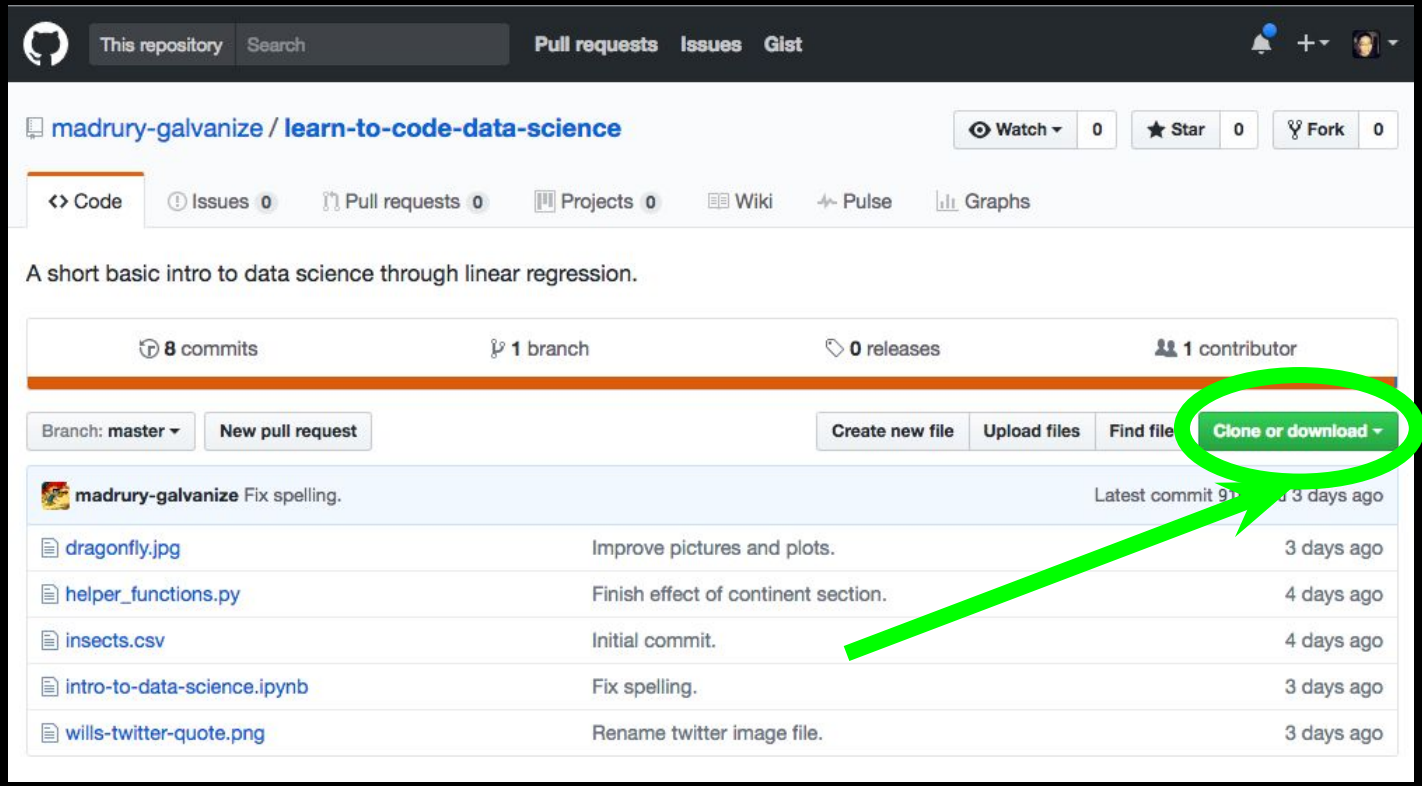
python.org/downloads

In case you need it, but Python is included in your Anaconda install.

2: Download the GitHub lesson

1. Go to: github.com/madrury-galvanize/learn-to-code-data-science/
   Or: bit.ly/madrury-ltc-ds
2. Clone or download the repo to your own computer (Remember where you put the files!)
   a. The key file for us: insects.csv

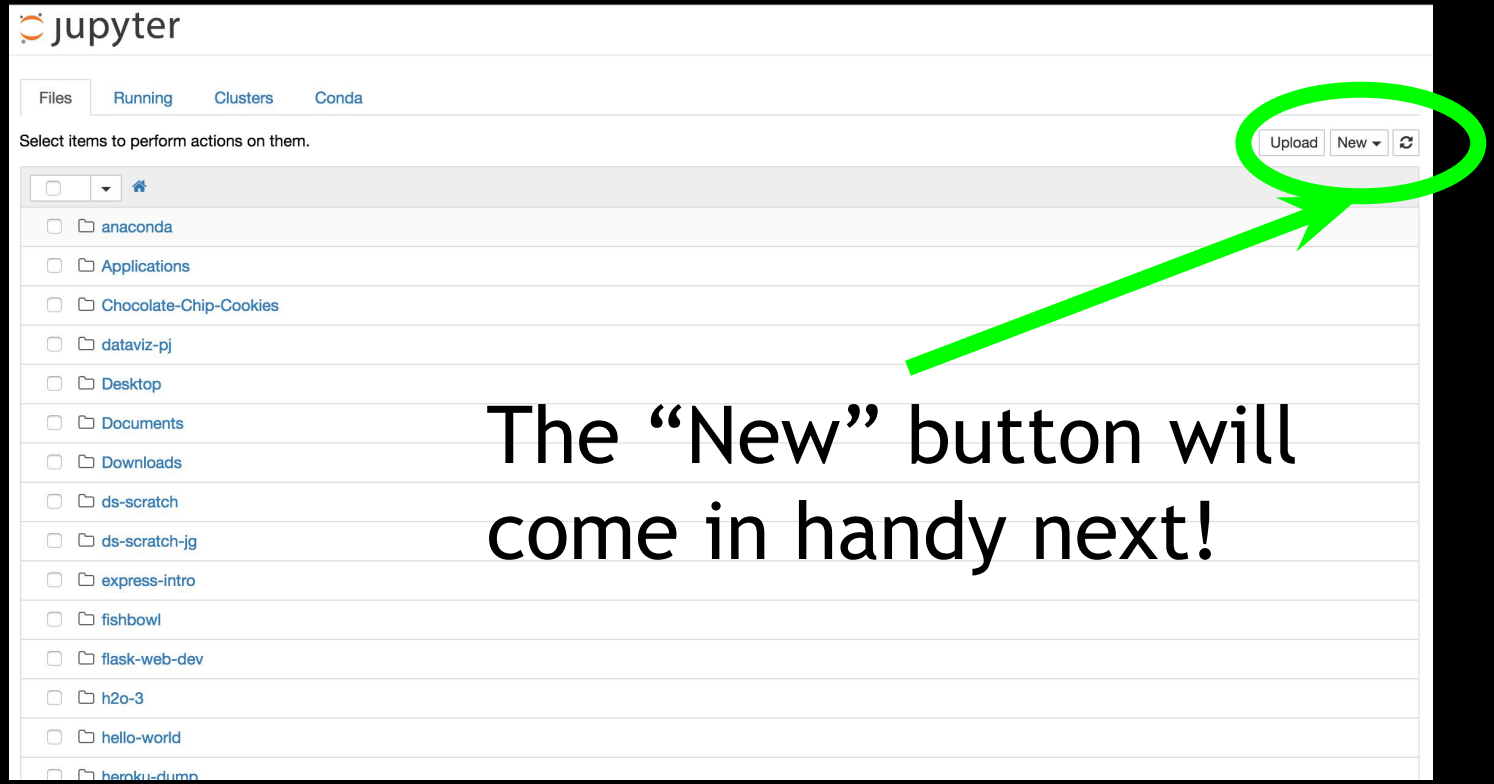
What you should see...



3. Let’s initialize Jupyter

1. In the terminal, navigate to your working directory where you saved the data files
2. Type “jupyter notebook” into the prompt
   Some computation should happen...
3. Go to your browser and type in this URL:
   http://localhost:8888/
   ^ (this may happen automatically)

What you should see...



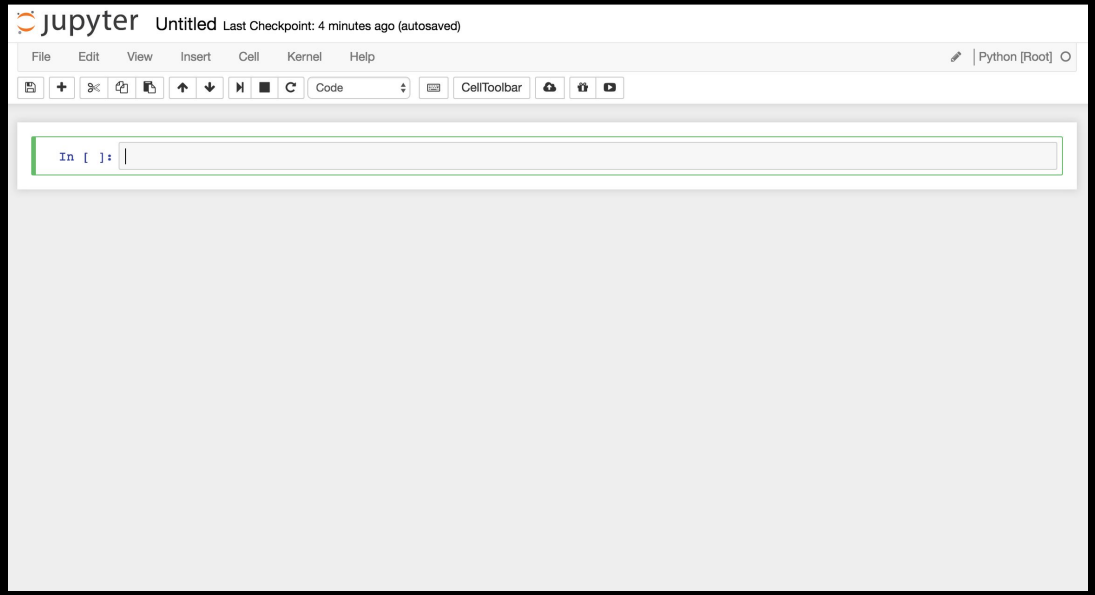
The “New” button will come in handy next!

Create a new Jupyter Notebook

5. Click on “New” in the top right corner
6. Select under “Notebooks” > “Python [root]” (or something similar)

Something should initialize immediately...

What you should see now...



If you see this, you are good to go!

If not, raise your hand!

Pictures of Pandas in Playgrounds

Setting up your computer can take time...

If you've done the following:

- ❑ Install Anaconda with Python 2.7 or higher
- ❑ Have a copy of the GitHub repo
- ❑ Initialized Jupyter Notebook

You're ready to move on to the next step!

In this course you will learn

- ❑ ~~Set up your computer for Jupyter Notebook~~
- ❑ Importing Libraries
- ❑ Loading and Inspecting Data
- ❑ Creating Visualizations
- ❑ Creating a Linear Regression

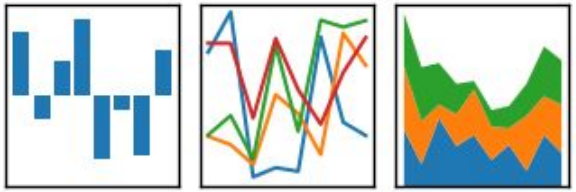
Importing Libraries

We’ll get buy with a little help from our friends

We're going to use the following:

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



NumPy

matplotlib

SM StatsModels
Statistics in Python

Let’s import what we need

```
import pandas as pd # load and manipulate data
import numpy as np # mathematical library
import statsmodel.formula.api as smf
# statistical analyses
from helper_functions import linear_model_summary
```

Let’s import what we need - matplotlib

```
%matplotlib inline
# Tells Jupyter to display the plots asap

import matplotlib.pyplot as plt
# matplotlib help us plot the data in his file
plt.style.use(‘ggplot’)
```

Let’s import what we need - rcParams

```
from pylab import rcParams
# No need to fuss with image sizes later on

rcParams[‘figure.figsize’] = 10, 6
```

If you've done the following:

- ❑ Wrote the code in Jupyter to import
  - ❑ Pandas
  - ❑ NumPy
  - ❑ Statsmodels
  - ❑ Helper_functions
  - ❑ Matplotlib
  - ❑ rcParams

You're ready to move on to the next step!

In this course you will learn

- ~~Set up your computer for Jupyter Notebook~~
- ~~Importing Libraries~~
- Loading and Inspecting Data
- Creating Visualizations
- Creating a Linear Regression

Inspecting Your Data

(Wait, what’s wrong with it?)

Let’s take a look into the data set!

```
>> !head ./insects.csv
^ # when you see “>>”, that’s our way
of saying we’d like you to type that
into your Notebook on a new line
```

What do you see?

Let’s import the data as a Python object

```
>> insects = pd.read_csv(‘./insects.csv’,
sep=‘\t’)
```

Let’s call ‘insects’ and see what happens.

```
>> insects
```

Did it work? Let's check!

	continent	latitude	wingsize	sex
0	1	35.5	901	0
1	1	37.0	896	0
2	1	38.6	906	0
3	1	40.7	907	0
4	1	40.9	898	0
5	1	42.4	893	0
6	1	45.0	913	0
7	1	46.8	915	0
8	1	48.8	927	0
9	1	49.8	924	0
10	1	50.8	930	0
11	0	36.4	905	0

Do you see 41 rows of data?

column headers:

- Continent
- Latitude
- Wingsize
- Sex

Get a description of the data:

```
>> insects.info()
```

If you've done the following:

- ❑ Explored your data's first 10 rows
- ❑ Loaded your data as a Python object
- ❑ Saw descriptive info about that object

You're ready to move on to the next step!

In this course you will learn

- ~~Set up your computer for Jupyter Notebook~~
- ~~Importing Libraries~~
- ~~Loading and Inspecting Data~~
- Creating Visualizations
- Creating a Linear Regression

Creating Visualizations

(Histograms)

Histogram

Let’s see a histogram of our data. Step 1!

```
>> column_names = {
    "continent": "Continent",
    "latitude": "Latitude",
    "wingsize": "Wing Span",
    "sex": "Sex"
}
```

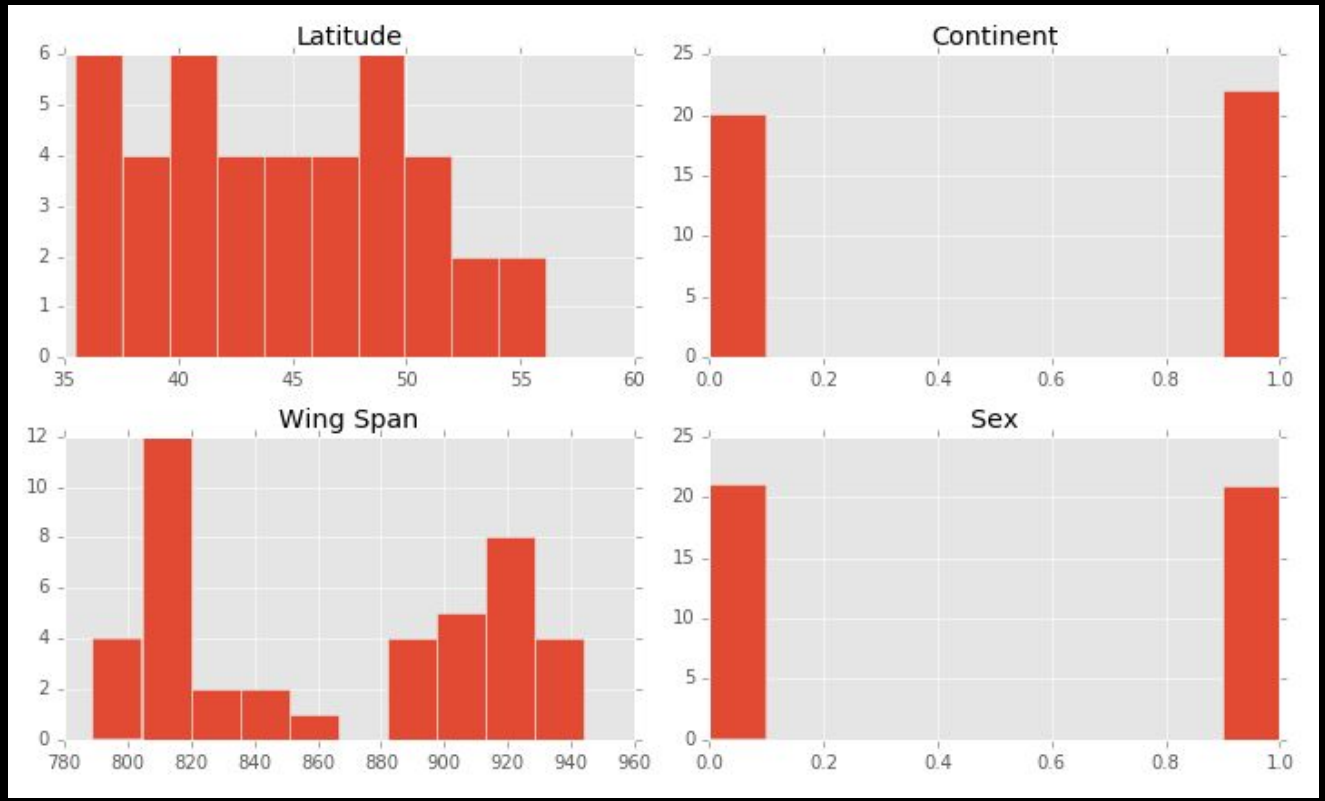
Histogram

Let’s see a histogram of our data. Step 2!

```
>> fig, axs = plt.subplots(2, 2)
for ax, (column, name) in zip(axs.flatten(),
column_names.iteritems()):
    ax.hist(insects[column])
    ax.set_title(name)

fig.tight_layout()
```

Here's what we should see!



Discussion

- Why do the data on the left look … different than that on the right?
  - Key concept: *binary/indicator values*
- What do you see happening with the data on wingspan?

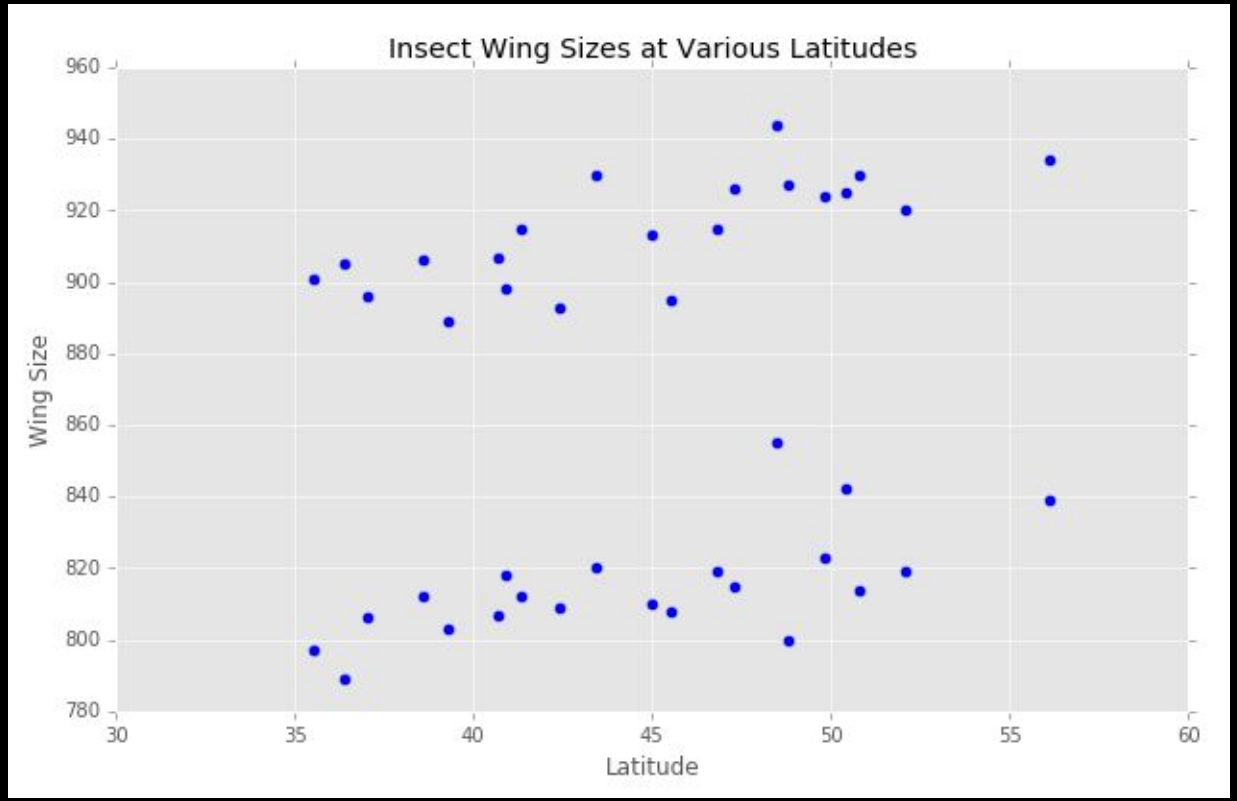
Creating Visualizations

(Scatterplots)

Scatterplots

```
fig, ax = plt.subplots()
ax.scatter(insects.latitude, insects.wingsize,
s=40)
ax.set_xlabel("Latitude")
ax.set_ylabel("Wing Size")
ax.set_title("Insect Wing Sizes at Various
Latitudes")
```

Here's what we should see!



Discussion

- What patterns do you see in the scatterplot?
- Can you form some hypothesis about the data?

Exploratory Data Analysis

(Let’s dig a little deeper!)

Explore the following questions

1. Are the two clusters associated with one of the other two variables in the dataset, continent or sex?
2. Is the increase of wing size as latitude increases real or illusory?
3. Does continent have any effect on wing size?
4. If the increase in wing size is real, does the *rate* of increase differ in the two clusters?

Let’s start with...

Are the two clusters associated with one of the other two variables in the dataset, continent or sex?

Here’s the code for ‘continent’

PART 1: Setting up the first plot

```
fig, ax = plt.subplots()
continent_boolean = insects.continent.astype(bool)

ax.scatter(
    insects.latitude[continent_boolean],
    insects.wingsize[continent_boolean],
    s=40, c="red", label="Continent 1")
```

Here’s the code for ‘continent’

Part 2: The second scatter plot

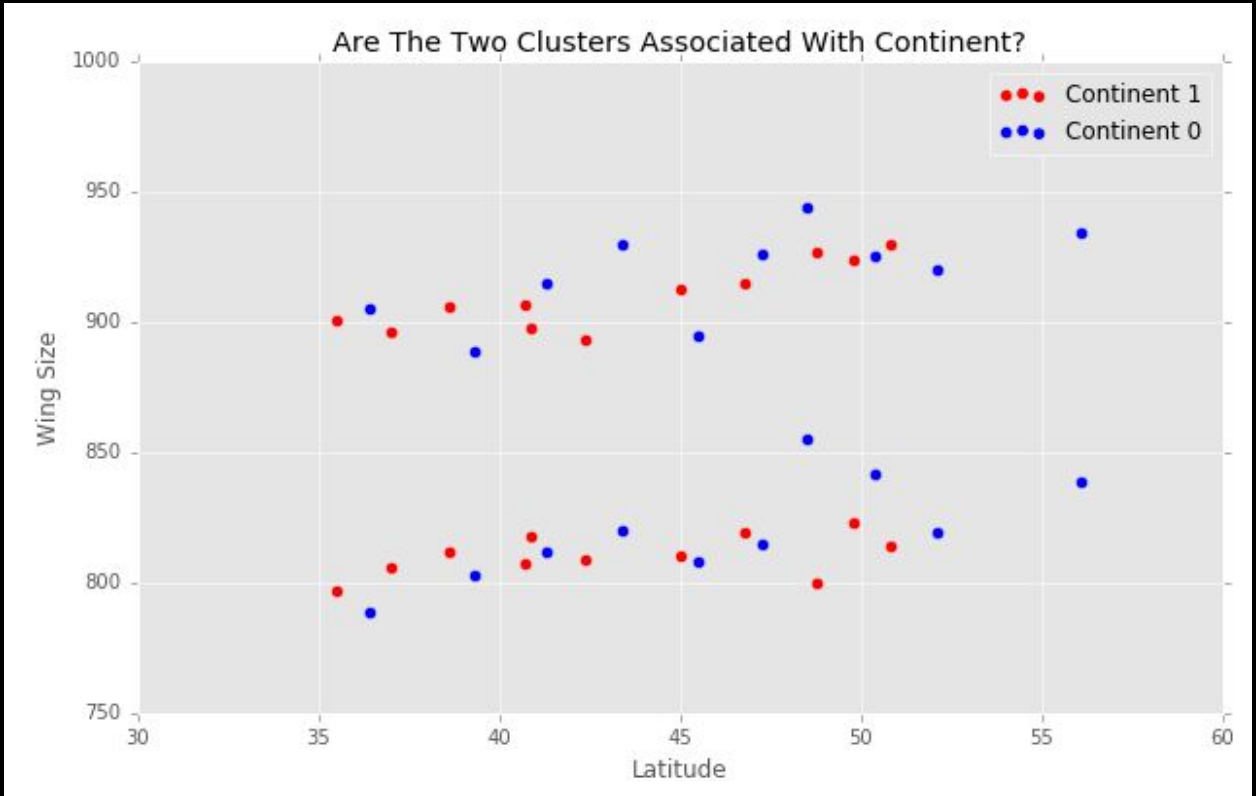
```
ax.scatter(insects.latitude[~continent_boolean],
           insects.wingsize[~continent_boolean],
           s=40, c="blue", label="Continent 0")
```

Here’s the code for ‘continent’

Part 3 (mostly for the visualization)

```
ax.set_xlabel("Latitude")
ax.set_ylabel("Wing Size")
ax.set_title("Are The Two Clusters Associated
With Continent?")
ax.legend()
```

Here's what we should see!



Discussion

- Do we see much of a difference when checking ‘continent’?
- What if we do the same for ‘sex’?

Here’s the code for ‘sex’

PART 1: Setting up the first plot

```
fig, ax = plt.subplots()
sex_boolean = insects.sex.astype(bool)

ax.scatter(
    insects.latitude[sex_boolean],
    insects.wingsize[sex_boolean],
    s=40, c="red", label="Male")
```

Here’s the code for ‘sex’

Part 2: The second scatter plot

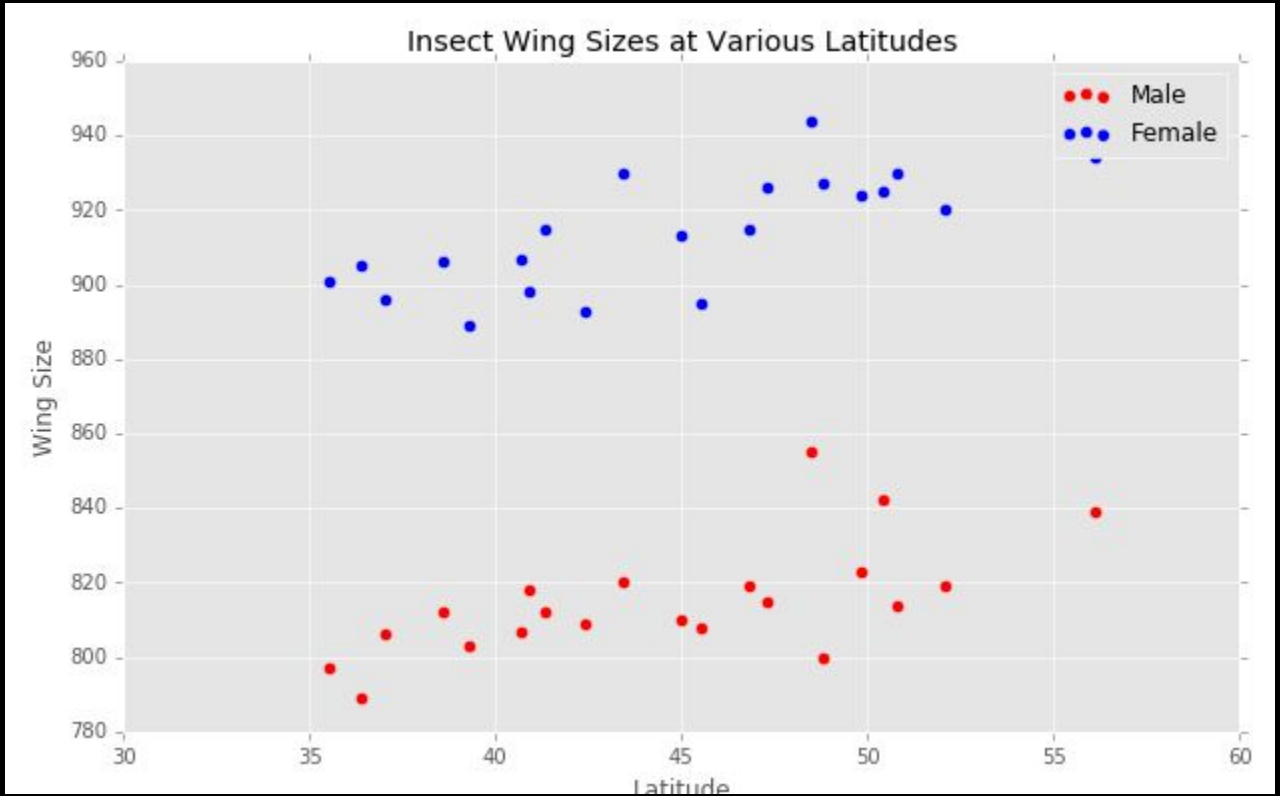
```
ax.scatter(insects.latitude[~sex_boolean],
           insects.wingsize[~sex_boolean],
           s=40, c="blue", label="Female")
```

Here’s the code for ‘sex’

Part 3 (mostly for the visualization)

```
ax.set_xlabel("Latitude")
ax.set_ylabel("Wing Size")
ax.set_title("Insect Wing Sizes at Various
Latitudes?")
ax.legend()
```

Here's what we should see!



Discussion

- Do we see much of a difference when checking ‘sex’?

If you've done the following:

- ❑ Created a scatterplot of 'continent'
- ❑ Created a scatterplot of 'sex'

You're ready to move on to the next step!

In this course you will learn

- ❑ ~~Set up your computer for Jupyter Notebook~~
- ❑ ~~Importing Libraries~~
- ❑ ~~Loading and Inspecting Data~~
- ❑ ~~Creating Visualizations~~
- ❑ Creating a Linear Regression

Linear Regression

(Talk nerdy to me!}

Try another question!

Is an increase in latitude associated with an increase in wing size?

$$\text{Wing Span} \approx a + b * \text{Latitude}$$

We'll need **linear regression**.

Here’s the code for ‘linear model’

```
linear_model = smf.ols(formula='wingsize ~ latitude',
data=insects)
insects_model = linear_model.fit()
linear_model_summary(insects_model)
```

Here's what we should see!

Linear Model Summary

Name	Parameter Estimate	Standard Error
Intercept	780.53	64.53
latitude	1.88	1.44

Wing Span ≈ a + b * Latitude

Let’s make a line according to ‘sex’

Step 1: (Re-use the code from earlier to make the sex scatterplots.)

Let’s make a line according to ‘sex’

Step 2: Here’s the code for a line graph.

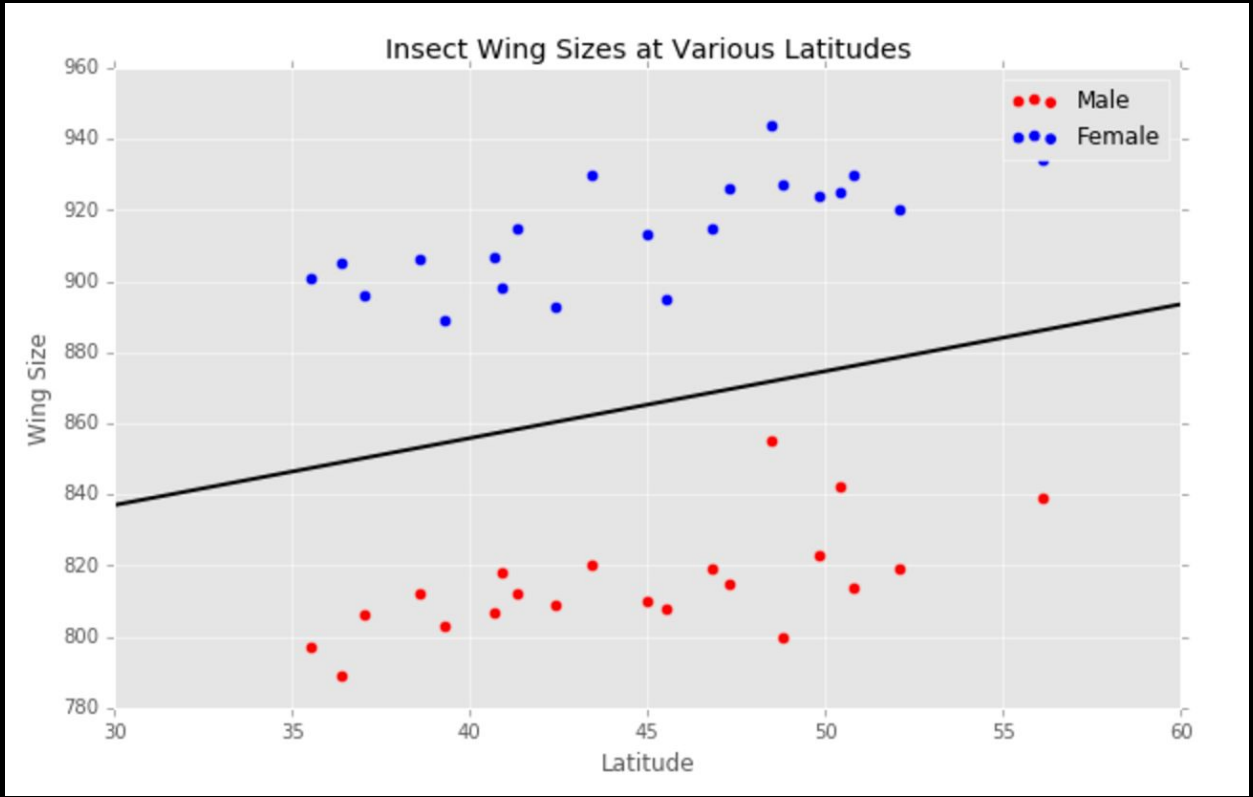
```
x = np.linspace(30, 60, num=250)
ax.plot(x, insects_model.params[0] +
           insects_model.params[1] * x,
        linewidth=2, c="black")
```

Let’s make a line according to ‘sex’

Step 2: Finish up the visualization.

```
ax.set_xlim(30, 60)
ax.set_xlabel("Latitude")
ax.set_ylabel("Wing Size")
ax.set_title("Insect Wing Sizes at Various
Latitudes")
ax.legend()
```

Here's what we should see!



Discussion

- We just made our first model! How well does it ‘fit’ our hypothesis?
- What else can we draw from this first attempt at a linear regression?

If you've done the following:

- ❑ Create a linear model of the data
- ❑ Create a visualization based on sex

You're ready to move on to the next step!

In this course you will learn

- ~~Set up your computer for Jupyter Notebook~~
- ~~Importing Libraries~~
- ~~Loading and Inspecting Data~~
- ~~Creating Visualizations~~
- ~~Creating a Linear Regression~~

Play around in the sandbox! Try to...

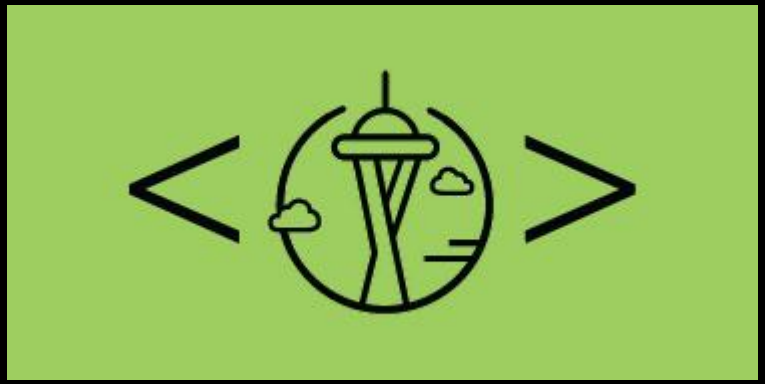
- Does continent have any effect on wing size?
- If the increase in wing size is real, does the *rate* of increase differ in the two clusters?

github.com/madrury-galvanize/learn-to-code-data-science
bit.ly/madrury-ltc-ds

You did it!

You are now a data scientist...ish.
Welcome to the cool kids club.

Keep the party going!



Come back for more!

Join our Meetups

Learn to Code Seattle

Seattle Data Science

Seattle Data Engineering

Startup Tech Seattle

Learn more on your own!

Go to: github.com/GalvanizeOpenSource/

Plenty of different courses available in learning to code!

Get yourself primed in data science

github.com/zipfian/data-science-primer

- Programming in Python
- Probability
- Statistics
- Linear Algebra
- SQL
- Machine Learning

Learn Data Science with Galvanize

Data Science Fundamentals: Intro to Python

- 6 week part-time workshop

Data Science Immersive Program

- 12 week full-time program

GalvanizeU

- 12 month program in San Francisco
- Fully-accredited by the University of New Haven

To learn more, visit galvanize.com/data-science
Or email enrollment@galvanize.com

powered by galvanize

Thank you for coming to galvanize

Email **Lee Ngo** at
lee.ngo@galvanize.com
or
Visit our website at
galvanize.com

This course has been brought to you by the evangelists of Galvanize.