# BIOL262 Lesson #11 – Data Analysis with Python and Pandas

GitHub Repository for Lesson: https://github.com/greenkidneybean/biol262
MyBinder Environment: https://bit.ly/2AnC81p

Data Science Workflow

1) What is your question
2) Identify appropriate dataset
3) Explore
4) Clean/Wrangle
5) Visualize

Definitions:
**Python** – high-level general-purpose programming language with a thriving community
**IDE** (integrated development environment) – software application that provides comprehensive facilities to develop software
**Jupyter** Notebooks – open-source web application that contains both live code and narrative text
**Pandas** – Python package used for data manipulation and analysis
**DataFrame** – a 2d labeled data structure, can be compared to an Excel spreadsheet

Homework:
For the Python/Pandas portion of this lesson a quiz is available on Canvas and must be submitted by Dec. 3rd at 5pm for credit. This repository contains a Jupyter notebook ('Python_Pandas_Homework.ipynb') to assist with the questions and can also be launched using the MyBinder link above.

Code Snippets:

```
# import pandas
import pandas as pd

# create dataframe
df = pd.DataFrame('data/ILINet.csv')

# functions to explore data
df.shape # returns the number of rows and columns
df.info() # returns information regarding column types and null values
df.describe() # returns general descriptive statistics for numeric columns
df.columns # returns the string labels for each column
df.head() # returns the top 5 rows of the dataframe
df.tail() # returns the last 5 rows of the dataframe
df.sample(10) # returns a random sampling of rows from the dataframe
```

```
# wrangling functions
df.isnull().sum() # combining 2 functions to get the count of null values in each column
df.replace() # replace a value in a column, or the name of a column
df['column_name'].astype('int') # changes the type of the column ('int', 'float','object')
df['column_name'].unique() # returns list of all unique values in column
df['column_name'].nunique() # returns number of unique values in column
df.pivot() # return a reshaped dataframe organized by provided index/column values

# saving a dataframe:
df.to_csv('file_name.csv') # save your dataframe as a comma separated file

# visualizing dataframes
# helpful packages and settings
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn')
%matplotlib inline # this is specific to jupyter notebooks

# simple plots using Pandas
df.plot()
df.hist()
df.boxplot()

# plots using Seaborn
plt = sns.swarmplot()
fig = fig.get_figure()
fig.save_fig('name_of_figure')
```