# KEY_Practice14_Intro_Stats_I

July 15, 2019

## 1 Practice with Statistics!

**Remember:** * The mean of an array can be calculated by `np.mean` and medians can be calculated with `np.mean`. * Means and medians are different types of central tendancy measures, which tell you about how the average of a dataset behaves.

First, import numpy and pandas:

[0]:
```
# load numpy and pandas

import numpy as np
import pandas as pd
```

[2]:
```
# mount Google Drive
from google.colab import drive
drive.mount('/content/gdrive')
path = '/content/gdrive/My Drive/SummerExperience-master/'
```

```
Drive already mounted at /content/gdrive; to attempt to forcibly remount, call
drive.mount("/content/gdrive", force_remount=True).
```

Load in the sample data from the Lesson:

[0]:
```
# load the csv file 'SampleData/iris.csv'
data_table = pd.read_csv(path + 'SampleData/iris.csv')
```

Now print out the table to see what other columns you can look at for practice:

[4]:
```
# Print the data_table array

data_table.head()
```

[4]:

| | sepal_length | sepal_width | petal_length | petal_width | species |
|---|---|---|---|---|---|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | setosa |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | setosa |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | setosa |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | setosa |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | setosa |

Pick one of the metrics other than `sepal_length` and subset this column into three different numpy arrays, one for each species ('setosa', 'versicolor', and 'virginica'):

[0]:
```
# Load one of the parameters above into three numpy arrays, based on species
# Hint, you can use the pandas query function to filter the arrays


petal_setosa = data_table.query('species == "setosa"')['petal_length']

petal_virginica = data_table.query('species == "virginica"')['petal_length']

petal_versicolor = data_table.query('species == "versicolor"')['petal_length']
```

Calculate the means of each of the species arrays:

[6]:
```
# Calculate the means of the three arrays you generated above

means = [
    np.mean(petal_setosa),
    np.mean(petal_virginica),
    np.mean(petal_versicolor)
]

means
```

[6]:
```
[1.4620000000000002, 5.552, 4.26]
```

Calculate the medians of each of the species arrays:

[7]:
```
# Calculate the medians of the three arrays you generated above:

medians = [
    np.median(petal_setosa),
    np.median(petal_virginica),
    np.median(petal_versicolor)
]

medians
```

[7]:
```
[1.5, 5.55, 4.35]
```

Compare the means and medians for each of the species, are the means and medians similar to each other? Do different species have different means?

[8]:
```
# Calculate the differences between the means you calculated above:

print(means[1] - means[0])
print(means[2] - means[1])
print(means[2] - means[0])
```

```
4.09
-1.2919999999999998
2.7979999999999996
```

What can you infer about each of the three different iris species based on this result?
Nice job! You just practiced:

- Calculating mean and medians using numpy
- Intepreting the results from these basic statistics