Basic Statistics in Python

When Should I use Each t-test?

- Single sample t-test
  - Comparing single value to a population
- Independent t-test
  - Comparing the means of two unrelated groups
  - IV = group
  - DV = continuous variable
- Dependent t-test
  - Comparing the means of two related groups
  - IV = time variable
  - DV = continuous variable

Assumption for t-tests

- Normality

- That means histograms!

dataFrame[‘columnName’].hist()

When Should I use the Others?

- Independent Chi-Square
  - Comparing frequencies of unrelated groups
  - IV = Categorical
  - DV = Categorical

- Correlation
  - Determining how related two variables are
  - Typically 2 continuous variables

Packages Important to Statistics

- pandas (for wrangling)
- scipy.stats (for the statistics info)
- matplotlib and/or seaborn (for graphing)

t-test Code

```
stats.ttest_1samp(dataFrame['columnName'], value)
```

```
ttest_ind(dataFrame.column[dataFrame.column == 'value'], dataFrame.column[dataFrame.column == 'second value'])
```

```
stats.ttest_rel()
```

Chi-Square Code

```
pd.crosstab(dataFrame['column'],
dataFrame['column2'])
```

```
stats.chi2_contingency(crosstabName)
```

Correlation Code

```
dataFrame['column1'].corr(dataFrame['column2'])

sns.heatmap(dataFrame.corr(), annot=True)
```