Python data science (continued)

Hamish Gibbs

Tutorial #3: String operations

- [pandas-cookbook: String operations (Chapter 6)](#)
  - This tutorial is about extracting information from text in pandas.
  - *Hint: Some of the most interesting information in tomorrow's dataset might be in string variables.*
- Core concepts:
  - Detecting keywords in strings:

```
1  weather_description.str.contains('Snow')
```

Tutorial #4: Data cleaning

- pandas-cookbook: Data cleaning (Chapter 7)

- Core concepts:

  - Detecting nan values stored as strings:

    ```
    1  na_values = ['NO CLUE', 'N/A', '0']
    2  requests = pd.read_csv(..., na_values=na_values, ...)
    ```

  - Altering DataFrame values in-place:

    ```
    1  zero_zips = requests['Incident Zip'] == '00000'
    2  requests.loc[zero_zips, 'Incident Zip'] = np.nan
    ```

Tutorial #5: sklearn - Linear Regression

- [sklearn - Linear Regression Example](#)

- Core concepts:

  - Fitting a model to data

    ```
    1  regr.fit(diabetes_X_train, diabetes_y_train)
    ```

  - Making predictions with a model

    ```
    1  regr.predict(diabetes_X_test)
    ```

  - Model evaluation

    ```
    1  mean_squared_error(diabetes_y_test, diabetes_y_pred)
    ```

Challenge scaffold

- challenge_scaffold.py
  - Look at the Challenge Guidelines and the project scaffold.
  - I will walk through the challenge scaffold!
  - Use this scaffold as a starting point for the challenge tomorrow.

Extra

- Start working with the dataset we will use tomorrow.

- Work on a few of the items in the challenge:

  - *What variables are in the dataset?*

  - *What are the data types of the variables?*

  - *Is there any missing data?*