pandas and dplyr functionality

Julian

June 1, 2015

pandas and dplyr functionality

On Day 2, we saw how to use the pandas Python module to import, summarize and manipulate data in a tabular format. The scripts accompanying the first hour of Day 3 demonstrated how to do the same with base R and the 'Hadleyverse' package dplyr . This slide set demonstrates similar functionality with each module/package.

Since importing data has already been covered, the focus will be on data manipulation.

Dimensionality in pandas

```
import pandas as pd
import numpy as np
density_url = 'http://www.census.gov/2010census/csv/pop_densit

density_data_2010 = pd.read_csv(density_url, skiprows = [0, 1,

density_data_dimen = density_data_2010.shape
density_data_cols = density_data_2010.columns.values.tolist()
density_data_idx = density_data_2010.index
pop_1910 = density_data_2010['1910_POPULATION'].values

print 'Data dimensions: %d rows, %d columns' % (density_data_d
print 'First 10 indices of population density DataFrame: ', de
```

```
## Data dimensions: 53 rows, 34 columns
## First 10 indices of population density DataFrame:  Int64Inc
```