UNSW SYDNEY | Australia's Global University

COMP9321 Data Services Engineering

Term1, 2021

Week 2: Exploring your Data in Pandas

What are Pandas DataStructures

- **Series**: A Series is a one-dimensional array-like object containing a sequence of values and an associated array of data labels, called its *index*. The simplest Series is formed from only an array of data.

Example:

```
myseries = pd.Series([4, 7, -5, 3])
myseries
0 4
1 7
2 -5
3 3
dtype: int64
```

What are Pandas DataStructures

DataFrame:A DataFrame represents a rectangular table of data and contains an ordered collection of columns, each of which can be a different value type (numeric, string, boolean, etc.). The DataFrame has both a row and column index;

Example:

```
data = {'state': ['Ohio', 'Ohio', 'Ohio', 'Nevada', 'Nevada', 'Nevada'],
'year': [2000, 2001, 2002, 2001, 2002, 2003],
'pop': [1.5, 1.7, 3.6, 2.4, 2.9, 3.2]}
frame = pd.DataFrame(data)
```

Understanding the Data (ask the right Questions)

- What is this dataset?
- What should I expect within this dataset?
- Basic concepts (e.g., domain knowledge)
- What are the questions that I need to answer?
- Does the dataset have some sort of a schema? (utilize domain knowledge)

Understanding the Data using Python

- You can use the describe() function to get a summary about the data excluding the NaN values. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.
- Use pandas .shape attribute to view the number of samples and features we're dealing with
- it’s also a good idea to take a closer look at the data itself. With the help of the head() and tail() functions of the Pandas library, you can easily check out the first and last 5 lines of your DataFrame, respectively.
- Use pandas .sample attribute to view a random number of samples from the dataset

Understanding your Data

```
>>> df = pd.read_csv(‘MyLovelyDataset.csv')
>>> df.head()                              #you can also use df.tail to get the last 5 rows

     Identifier        Type of Company        Location
0        206                 NaN               Boston
1        216                 Law               London; Virtue & Yorston
2        218                 n/a               Sydney
3        472                 Finance           London
4        480                 Health            NY
```

*http://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/

Understanding your Data (Cont’d)

- If you have many columns and you want to understand what you have

```
>>> df = pd.read_csv(‘MyLovelyDataset.csv')
>>> list(df)                                        # gets list of column names
```

```
[‘Identifier’, ‘Type of Company’, ‘Location’]
```

*http://www.developintelligence.com/blog/2017/08/data-cleaning-pandas-python/

Useful Resource

- Book: Python for Data Analysis, Second Edition, Wes McKinney