ENGR 1330: Computational Thinking with Data Science

Lesson 9: Pandas In Python

Dinesh S. Devarajan
Whitacre College of Engineering
Texas Tech University

Topic Outline

- Pandas library
  - ✓ Data representation: Dataframes
  - ✓ Data operations: Indexing, summarizing statistics filling and dropping values, and read/write files

Objectives

- To be able to represent data in the form of dataframes via the Pandas library
- To be able to access and manipulate data within a dataframe
- To be able to obtain basic statistical measures of data within a dataframe

Computational Thinking Concepts

Pandas dataframes $\Rightarrow$ **Data representation**

Data interpretation, manipulation, and analysis of Pandas dataframes $\Rightarrow$ **Decomposition**

Pandas in Python

Pandas

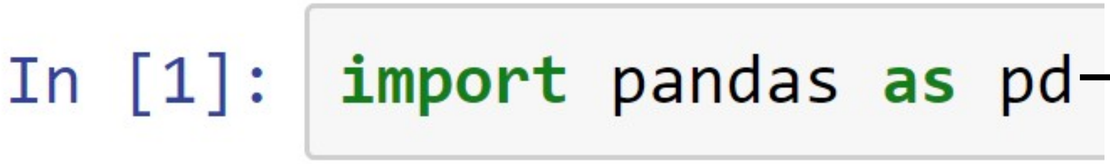
- Pandas: Derived from the term ‘Panel Data’
- Primary data structure is dataframe
- Dataframe: 2-dimensional mutable and heterogenous tabular data structure
- Popular among statisticians and data scientists

Features of Pandas

- Features:
  - ✓ Provides rich data structures and functions designed to make working with data fast, easy, and expressive
  - ✓ Useful in data manipulation, cleaning, and analysis
  - ✓ Excels in performance and productivity

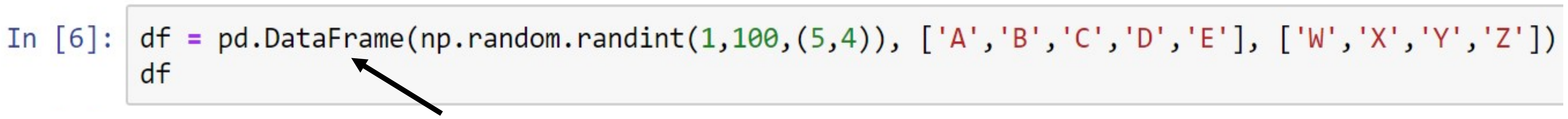
Pandas Dataframes

- Creating a dataframe:



```
In [1]: import pandas as pd
```

→ Importing Pandas library



```
In [6]: df = pd.DataFrame(np.random.randint(1,100,(5,4)), ['A','B','C','D','E'], ['W','X','Y','Z'])
        df
```

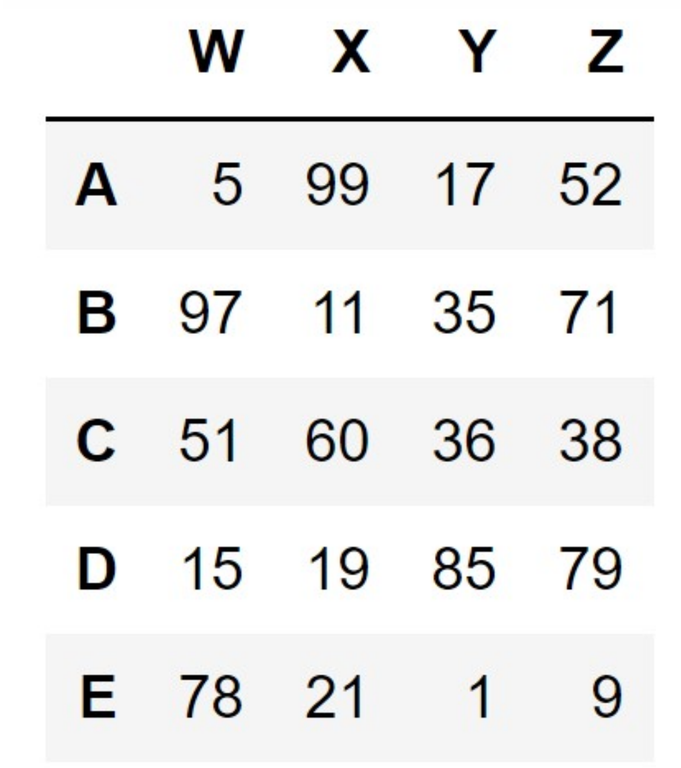
Function to create a Pandas dataframe

- What will be the shape of the above 2D Pandas dataframe?

(Demo)

Dataframes: Indexing

df =



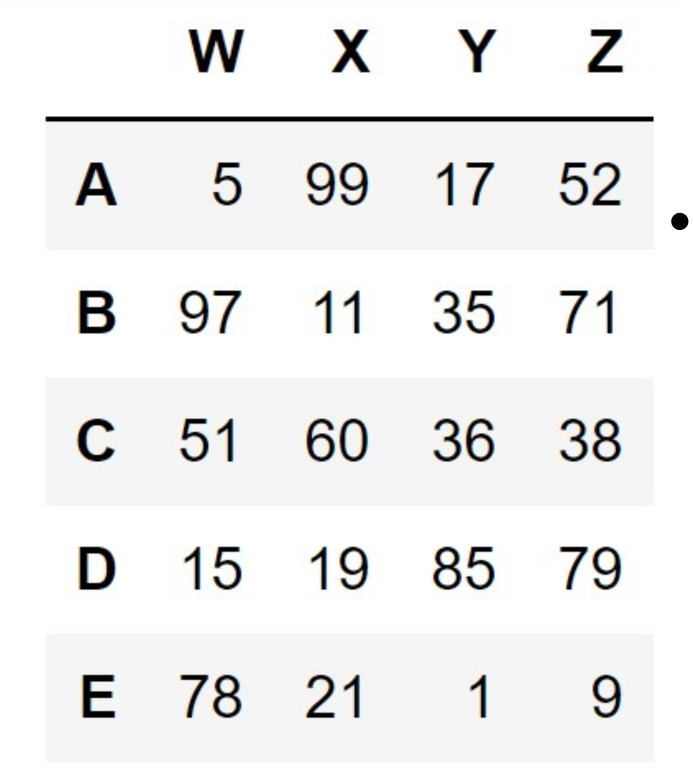
	W	X	Y	Z
A	5	99	17	52
B	97	11	35	71
C	51	60	36	38
D	15	19	85	79
E	78	21	1	9

- How would you index and slice all the elements of column ‘X’ in the above dataframe named ‘df’?

(Demo)

Dataframes: Indexing

df =



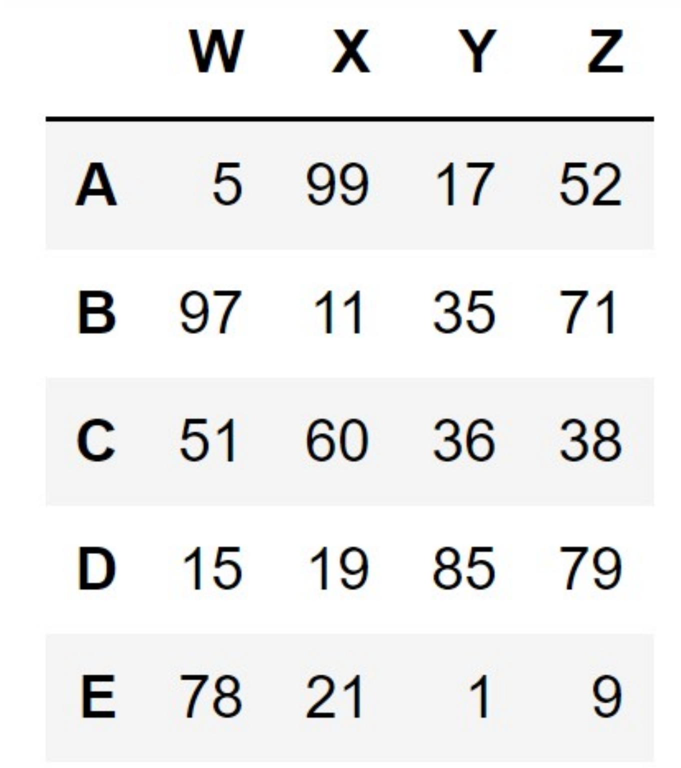
	W	X	Y	Z
A	5	99	17	52
B	97	11	35	71
C	51	60	36	38
D	15	19	85	79
E	78	21	1	9

- How would you index and slice all the elements of columns ‘X’ and ‘Z’ in the above dataframe named ‘df’?

(Demo)

Dataframes: Indexing

df =



	W	X	Y	Z
A	5	99	17	52
B	97	11	35	71
C	51	60	36	38
D	15	19	85	79
E	78	21	1	9

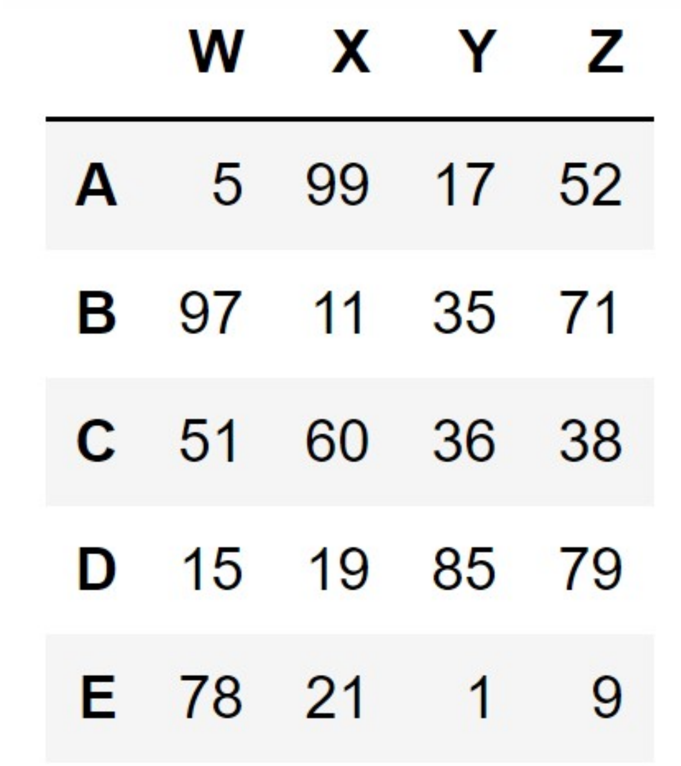
- How would you index and slice all the elements of row ‘C’ in the above dataframe named ‘df’?

- When dealing with row indexing, use loc[ ] indexer

(Demo)

Dataframes: Indexing

df =



	W	X	Y	Z
A	5	99	17	52
B	97	11	35	71
C	51	60	36	38
D	15	19	85	79
E	78	21	1	9

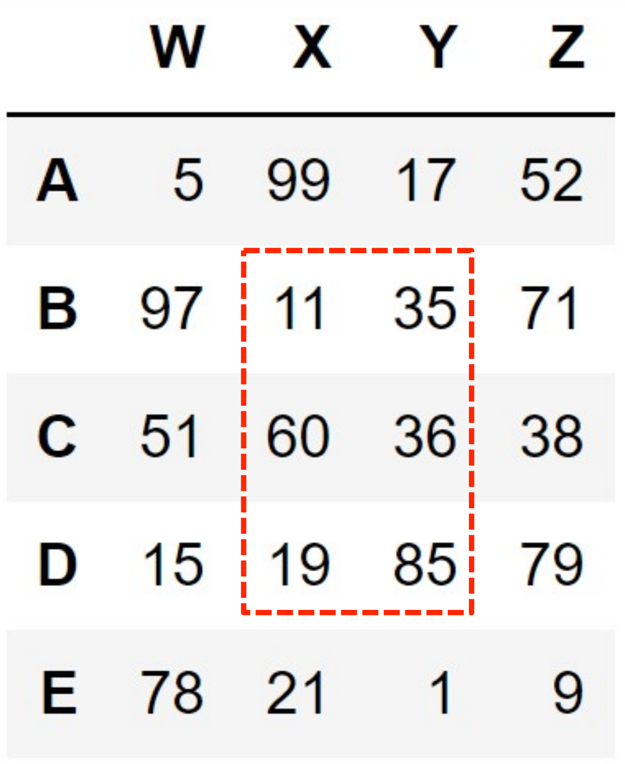
- How would you index and slice all the elements of rows ‘C’ and ‘E’ in the above dataframe named ‘df’?

- When dealing with row indexing, use loc[ ] indexer

(Demo)

Dataframes: Indexing

df =



	W	X	Y	Z
A	5	99	17	52
B	97	11	35	71
C	51	60	36	38
D	15	19	85	79
E	78	21	1	9

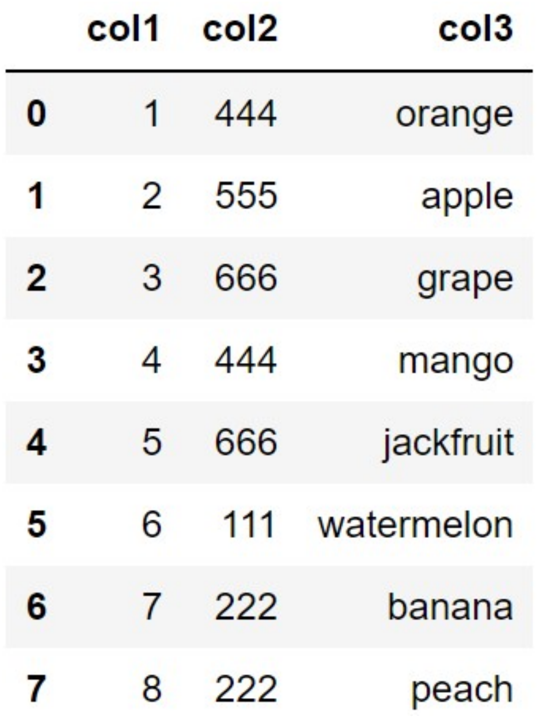
- How would you index and slice the elements within the red-dashed box from the dataframe named ‘df’?

- When dealing with row indexing, use loc[ ] indexer

(Demo)

Dataframes: Conditional Selection

df =



	col1	col2	col3
0	1	444	orange
1	2	555	apple
2	3	666	grape
3	4	444	mango
4	5	666	jackfruit
5	6	111	watermelon
6	7	222	banana
7	8	222	peach

- What fruit corresponds to the number 555 in ‘col2’?
- What fruit corresponds to the minimum number in ‘col2’?

(Demo)

Dataframes: Basic Operations

- Functions to do basic operations on Pandas dataframes

  - ✓ head( ): Returns first 5 rows of a dataframe
  - ✓ info( ): Returns information such as number of rows and columns about a dataframe
  - ✓ describe( ): Returns basic statistical measures of a dataframe

(Demo)

Dataframes: Basic Operations

- Functions to do basic operations on Pandas dataframes

  ✓ sum( ): Returns the sum of a column or a row

  ✓ unique( ): Returns the unique elements in a column

  ✓ nunique( ): Returns the number of unique elements in a column

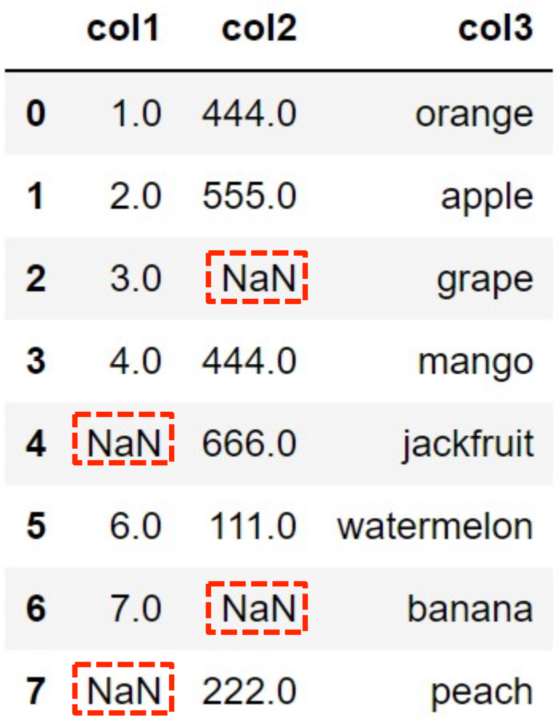
  ✓ value_counts( ): Returns the number of occurrences of each unique value

(Demo)

Dataframes: Dropping values

- Often, the data will consist of missing values ‘NaN’

df =

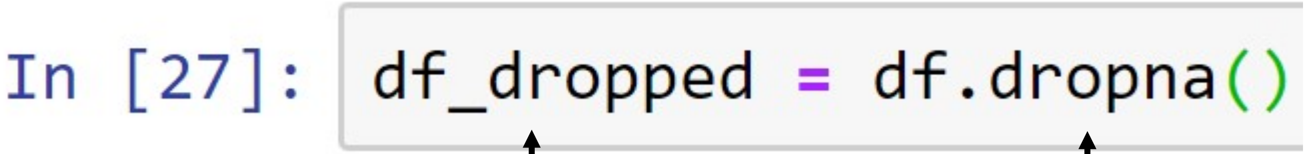


	col1	col2	col3
0	1.0	444.0	orange
1	2.0	555.0	apple
2	3.0	NaN	grape
3	4.0	444.0	mango
4	NaN	666.0	jackfruit
5	6.0	111.0	watermelon
6	7.0	NaN	banana
7	NaN	222.0	peach

- Missing values lead to problems in the data analysis process

Dataframes: Dropping values

- You can use the dropna( ) function to drop all the rows consisting of the missing values



```
In [27]: df_dropped = df.dropna()
```

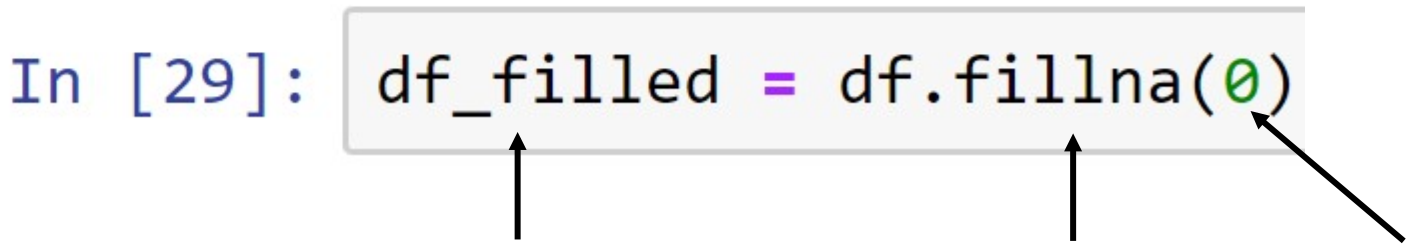
New dataframe after filling values

Function to drop values

(Demo)

Dataframes: Filling values

- You can also use the fillna( ) function to fill values (e.g. a value of '0' in the place of 'NaN') in the place of missing values

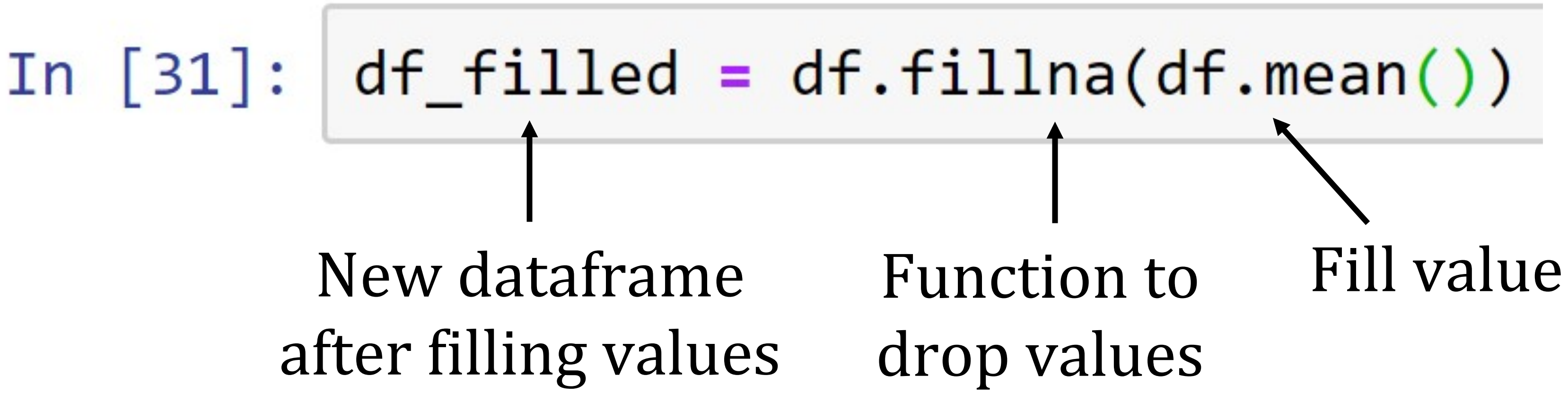


```
In [29]: df_filled = df.fillna(0)
```

(Demo)

Dataframes: Filling values

- You can also use the fillna( ) function to fill values (e.g. mean value of each column in the place of ‘NaN’) in the place of missing values

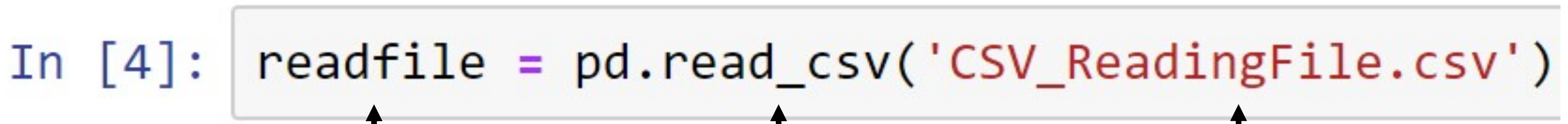


```
In [31]: df_filled = df.fillna(df.mean())
```

(Demo)

Dataframes: Reading a File

- Objective is to read the data in a ‘.csv’ (comma separated values) file and print it as a dataframe



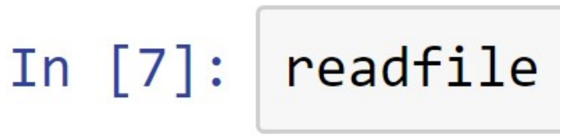
```
In [4]: readfile = pd.read_csv('CSV_ReadingFile.csv')
```

File object

Function to read .csv file

File name

- Printing the contents of the .csv file to the output screen

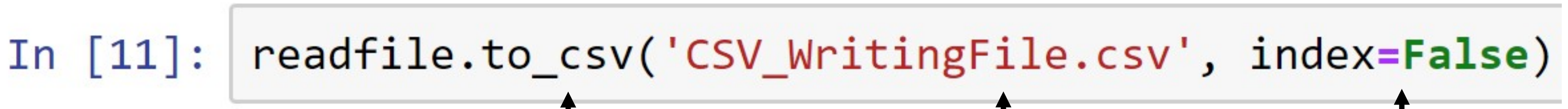


```
In [7]: readfile
```

(Demo)

Dataframes: Writing a File

- Objective is to write the data in a new '.csv' (comma separated values) file



```
In [11]: readfile.to_csv('CSV_WritingFile.csv', index=False)
```

Function to read .csv file

File name

Excludes row labels

- Note: File name that you give will first be created in the same folder where the Jupyter notebook is present

(Demo)

Summary

- Concepts of representing data in the form of Pandas dataframes are covered

- Concepts of interpreting, manipulating, and analyzing data within Pandas dataframes are covered