PYTHON FOR DATA SCIENCE

CHEAT SHEET

Python Pandas

What is Pandas?

It is a library that provides easy to use data structure and data analysis tool for Python Programming Language.

Import Convention

import pandas as pd – Import pandas

Pandas Data Structure

- Series:

```
s = pd.Series([1, 2, 3, 4], index=['a', 'b', 'c', 'd'])
```

- Data Frame:

```
data_mobile = {'Mobile': ['iPhone', 'Samsung',
'Redmi'], 'Color': ['Red', 'White', 'Black'], 'Price': [High,
Medium,Low]}
df = pd.DataFrame(data_mobile,
columns=['Mobile', 'Color', 'Price'])
```

Importing Data

- pd.read_csv(filename)
- pd.read_table(filename)
- pd.read_excel(filename)
- pd.read_sql(query, connection_object)
- pd.read_json(json_string)

Exporting Data

- df.to_csv(filename)
- df.to_excel(filename)
- df.to_sql(table_name, connection_object)
- df.to_json(filename)

Create Test / Fake Data

- pd.DataFrame(np.random.rand(4,3)) - 3 columns and 4 rows of random floats
- pd.Series(new_series) - Creates a series from an iterable new_series

Plotting

- Histogram: df.plot.hist()
- Scatter Plot: df.plot.scatter(x='column1',y='column2')

Operations

View DataFrame Contents:

- df.head(n) - look at first n rows of the DataFrame.
- df.tail(n) – look at last n rows of the DataFrame.
- df.shape() - Gives the number of rows and columns.
- df.info() - Information of Index, Datatype and Memory.
- df.describe() - Summary statistics for numerical columns.

Selection:

- iloc
  - df.iloc[0] - Select first row of data frame
  - df.iloc[1] - Select second row of data frame
  - df.iloc[-1] - Select last row of data frame
  - df.iloc[:,0] - Select first column of data frame
  - df.iloc[:,1] - Select second column of data frame
- loc
  - df.loc([0], [column labels]) - Select single value by row position & column labels
  - df.loc['row1':'row3', 'column1':'column3'] - Select and slicing on labels

Sort:

- df.sort_index() - Sorts by labels along an axis
- df.sort_values⊠by='Column label'⊠ - Sorts by the values along an axis
- df.sort_values(column1) - Sorts values by column1 in ascending order
- df.sort_values(column2,ascending=False) - Sorts values by column2 in descending order from one column
- df.groupby([column1,column2]) - Returns a groupby object values from multiple columns
- df.groupby(column1)[column2].mean() - Returns the mean of the values in column2, grouped by the values in column1
- df.groupby(column1)[column2].median() - Returns the mean of the values in column2, grouped by the values in column1

Functions

Mean:

- df.mean() - mean of all columns Median
- df.median() - median of each column Standard Deviation
- df.std() - standard deviation of each column

Max

- df.max() - highest value in each column Min
- df.min() - lowest value in each column Count
- df.count() - number of non-null values in each DataFrame column

Describe

- df.describe() - Summary statistics for numerical columns

FURTHERMORE:

Python for Data Science Certification Training Course