PANDAS

**Data Science** or **data analytics** is a process of analyzing large set of data points to get answers on questions related to that data.

**Pandas** is a python module that makes data science easy and effective

Process of cleaning messy data is called **data mugging** or **data wrangling**.

We can do data analytics in excel but if data grows in size then it will be difficult to handle it and it will be slow as well, we can do it using python program but in that case the code will be too lengthy so the best way for data analytics is ‘**Pandas’** module

**from** **pandas** **import** DataFrame, read\_csv

**import** **matplotlib.pyplot** **as** **plt**

**import** **pandas** **as** **pd** *#this is how I usually import pandas*

**import** **sys** *#only needed to determine Python version number*

**import** **matplotlib** *#only needed to determine Matplotlib version number*

*# Enable inline plotting*

%**matplotlib** inline

.

Create Data

names = ['Bob','Jessica','Mary','John','Mel']

births = [968, 155, 77, 578, 973]

BabyDataSet = list(zip(names,births))

BabyDataSet

We now will use the **pandas** library to export this data set into a csv file.

**df** will be a **DataFrame** object. You can think of this object holding the contents of the BabyDataSet in a format similar to a sql table or an excel spreadsheet.

df = pd.DataFrame(data = BabyDataSet, columns=['Names', 'Births'])

df

**Export to csv file:**

df.to\_csv('births1880.csv',index=**False**,header=**False**)

Get data

Location = r'C:\Users\david\notebooks\update\births1880.csv'

df = pd.read\_csv(Location)

df = pd.read\_csv(Location, names=['Names','Births'])

* **Delete the csv file:**

**import** **os**

os.remove(Location)

* To check the data types of the columns:

df.types

df.columnName.dtype

Analyze data

* Sort value by births:

Sorted = df.sort\_values(['Births'], ascending=**False**)

* Find max of birth value:

df['Births'].max()

**Notes:**

1. ***df['Names']*** - This is the entire list of baby names, the entire Names column
2. ***df['Births']***- This is the entire list of Births in the year 1880, the entire Births column
3. ***df['Births'].max()*** - This is the maximum value found in the Births column
4. **[df['Births'] == df['Births'].max()]** **IS EQUAL TO** [Find all of the records in the Births column where it is equal to 973]
5. **df['Names'][df['Births'] == df['Births'].max()]** **IS EQUAL TO** Select all of the records in the Names column **WHERE** [The Births column is equal to 973]
6. **create graph:**

df['Births'].plot()

1. **random\_names** = Select a random name from the name list and do this n times

random.seed(500)

random\_names = [names[random.randint(low=0,high=len(names))] **for** i **in** range(1000)]

1. **Generate a random numbers between 0 and 1000:**

births = [random.randint(low=0,high=1000) **for** i **in** range(1000)]

births[:10]

1. df.info() : to get he info about the dataframe we created
2. df.head() and df.tail() : to get first and last 5 records from df
3. df['Names'].unique() : to find unique records
4. *# Create graph*

df['Births'].plot.bar()

Key Features of Pandas

1. Fast and efficient DataFrame object with default and customized indexing.
2. Tools for loading data into in-memory data objects from different file formats.
3. Data alignment and integrated handling of missing data.
4. Reshaping and pivoting of date sets.
5. Label-based slicing, indexing and subsetting of large data sets.
6. Columns from a data structure can be deleted or inserted.
7. Group by data for aggregation and transformations.
8. High performance merging and joining of data.
9. Time Series functionality.

Introduction to Data Structures

Pandas deals with the following three data structures −

* **Series**
* **DataFrame**
* **Panel**

These data structures are built on top of Numpy array, which means they are fast.

### Mutability

All Pandas data structures are value mutable (can be changed) and except Series all are size mutable. Series is size immutable.

There are a lot in pandas so I can’t create notes for all of them, please refer below link and read online!!! ☺

References: <https://www.tutorialspoint.com/python_pandas/python_pandas_series.htm>

<https://pandas.pydata.org/pandas-docs/stable/tutorials.html>