**Collab experiments with Data Frames using Python and Pandas**

Some data:

https://resources.oreilly.com/examples/0636920023784/-/tree/master/pydata-book-master

Reference book: Python for Data Analysis, Wes McKinney (2013)

Use Google Colab, Google drive

Data: main

# colab

# textbook: Python for Data Analysis

# pages 4, 5

!pip install pandas

!pip install matplotlib

import pandas as pd

**Grouping data frames**

# Create a folder in google drive named ‘bigdata’

# Upload main-sales-data.csv to the above folder

Accessing the file in Google Collab:

from google.colab import drive

drive.mount('/content/gdrive')

!ls '/content/gdrive/My Drive/bigdata/main-sales-data.csv'

x = pd.read\_csv('/content/gdrive/My Drive/bigdata\_course/main-sales-data.csv')

# Page 262, (Diagram 252)

# Group by Country

g\_country = x.groupby(['Country'])

m\_sum = g\_country['Sales Pound']

m\_sum.agg('mean')

m\_sum.agg('sum')

m\_sum.agg(['sum', 'std', ‘peak\_to\_peak’])

m\_sum.agg(['sum', 'std'])

**Merging two data sets**

**(Page 177)**

main-sales, order-salesperson, names\_salesperson (All csv files)

y = pd.read\_csv('/content/gdrive/My Drive/bigdata\_course/order-salesperson.csv')

merge1 = pd.merge(x, y, on='Order ID')

z = pd.read\_csv('/content/gdrive/My Drive/bigdata\_course/names-salesperson.csv')

merge2 = pd.merge(merge1, z, left\_on='SalesPersonID', right\_on='ID')

g\_name = merge2.groupby('Name')

m\_qty = g\_name['Sales Pound']

m\_qty.agg('sum')

**Plotting**

(Page 219)

import matplotlib.pyplot as plt

# ax1 = fig.add\_subplot(2, 2, 1)

# Above: 2x2 subplots. We are adding to the number 1

sm = m\_qty.agg('sum')

ax = sm.plot(kind='bar')

ax.set\_title('Sales Performance by Sales Person')