To EPG Tec Blog Website – last updated 11/28/17..1/29/18…3/16/18…4/11/18

For Accountants who deal with big data set dump from the IS department

To analyze huge data set with endless rows and columns from the data dump, how to use Python to extract the data and perform the meaningful analysis? How to visualize the financial analysis?

The example illustrated is in csv or xlsx format since most likely that IS department would only provide the output data in either format.

Topic:

* Use Python to Import dataset in CSV or Excel data (first blog)
* To add and perform math average in Python for data validation
* After Reading Excel in Python, how to get the clean data for IS report spec?
* Put data in SQL Lite format for data analysis
* use Visio for mockup presentation purpose for reporting purpose.
* How to mockup the report so IS can understand the report spec?

**(A) Use Python to Import dataset in CSV or Excel format**

I use 2 sample data files,

* First sample data is 2015 USA census with 20,000 rows and 150 columns
* Second sample data is 10 stocks historical daily price from 1990-2011

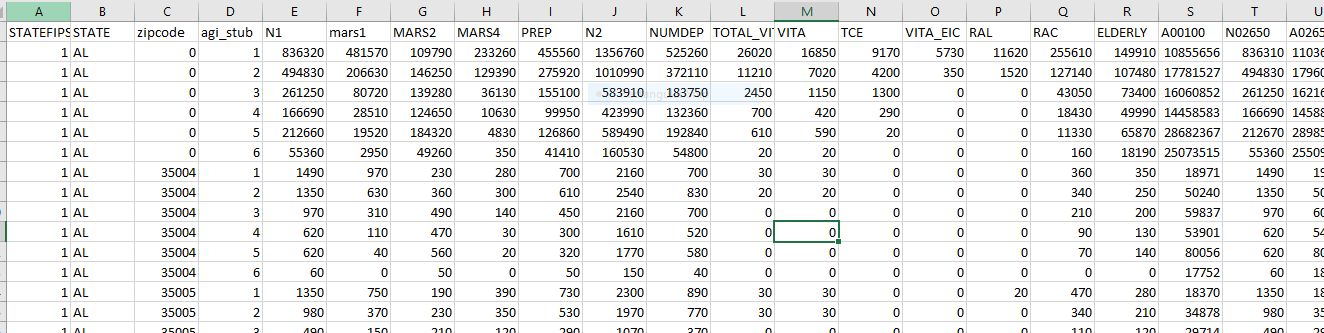
1. How to use Python to read Excel file?

For example, you get a huge dataset from IS department, when you glance the excel file, your head spin. The file has endless rows and columns and you need to analysis the data in the timely basis.

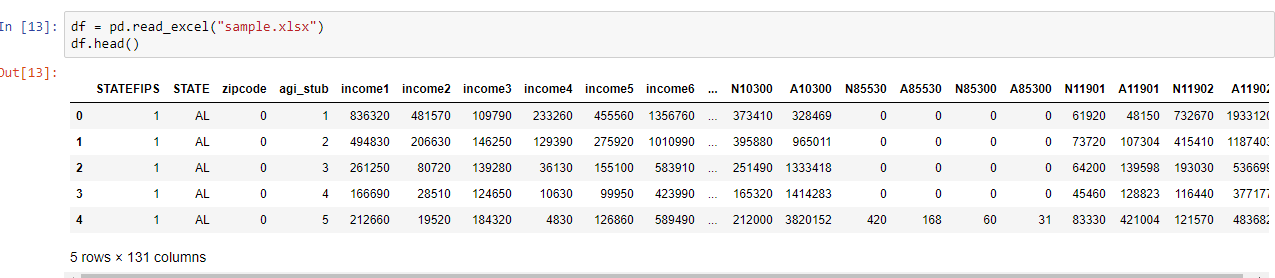
Step 1:

Say the excel file name called “sample.xlsx” with 20,000 rows and 150 columns.

This sample data is from gov.org - 2015 census data by zip code. For illustration purpose, the excel file saved the first 1,000 rows and 150 columns.



In python: to read the first 5 rows of the excel data



Step 2: An interesting note – there are several ways to read Excel file from Python Pandas.

pd = pd.read\_excel(“sample.xlsx”)

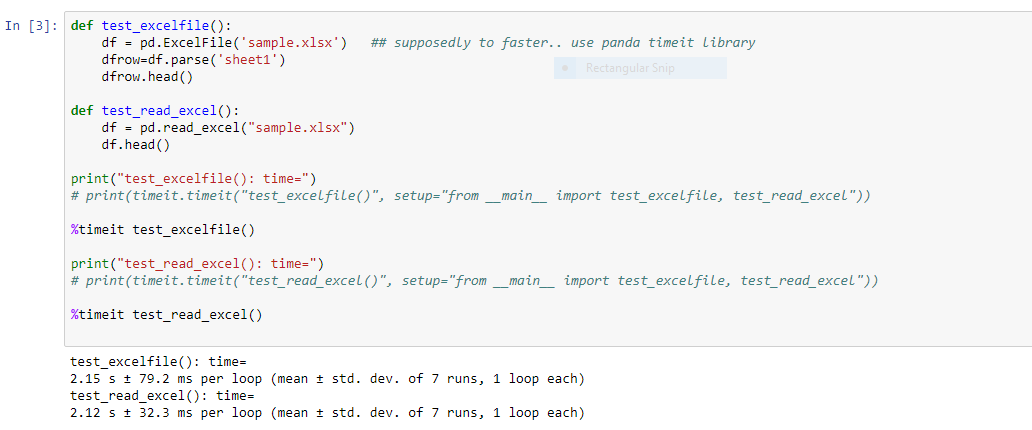
or

pd = pd.ExcelFile(‘sample.xlsx’)

Both commands can read Excel file.

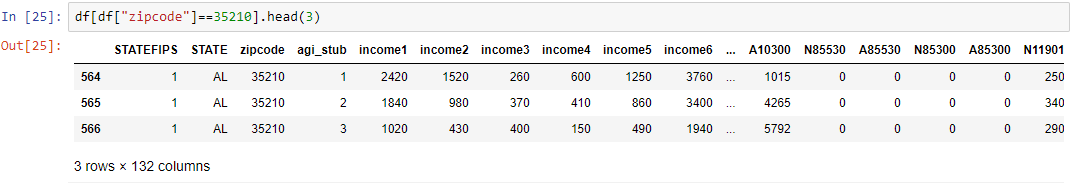
Which command should you use?

* You can use Pandas Timeit to time how long does it take to run the reports
* General speaking, read\_excel commands is faster



Step 3: How to filter data in Python, quick and save time instead of Excel? If you need to see the first three rows for zip code 35210:

Use command df[df[‘zipcode’]==35210.head(3)

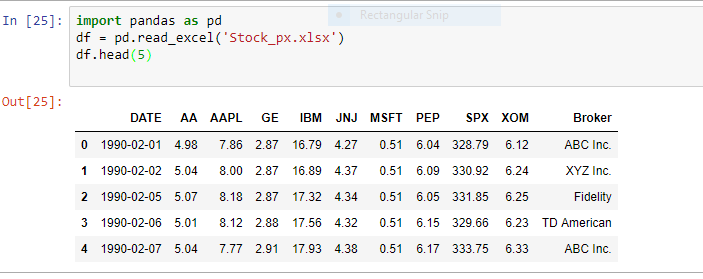


10 stocks historical price from 1990-2011

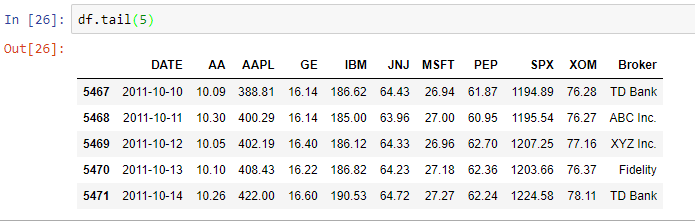
Step 1: Import pandas and read the Excel file (specified the directory)

This sample file is stock market value file with 100 stocks.

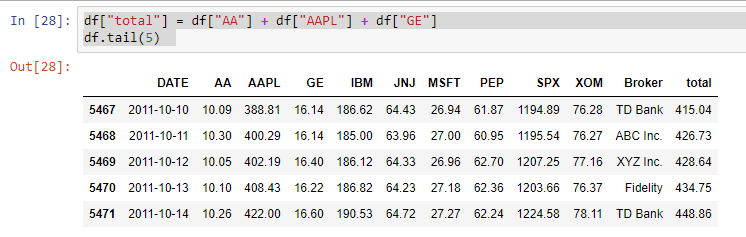
To view the first 5 rows of data



To view last five rows of data



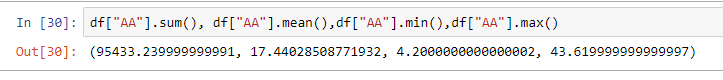
Step 2: Sum up stock price of AA, AAPL and GE



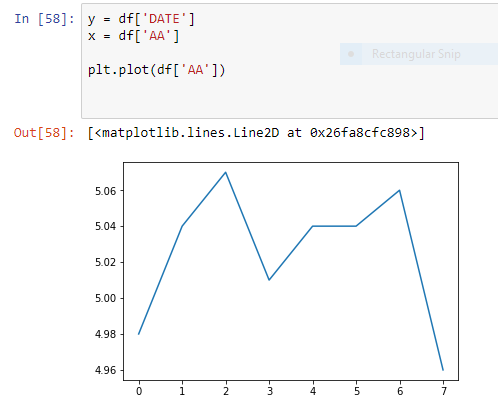
Step 3: data analysis – it’s easy to get the min, max in Python

The minimum stock price of Apple between 2/1/1990 and 10/14/2011 was $4.2

The maximum stock price of Apple between 2/1/1990 and 10/14/2011 was $43.619



Step 4: import matplot pyplot to get the chart



Step 5 - set index and only select specific columns , report only want to select DATE and stock AA

