By the end of this activity, you will be able to:

1. Read a CSV file into a Pandas DataFrame
2. View the contents and shape of a DataFrame
3. Filter rows and columns of a DataFrame
4. Calculate the average and sum of a column in a DataFrame
5. Combine two DataFrames by joining on a single column

This activity consists of programming in a Jupyter Python Notebook. If you have not already started the Jupyter server, follow the instructions in the Reading entitled *Starting Jupyter for Python Notebooks.*

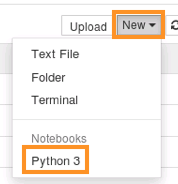
Step 1. **Open a web browser and create a new Jupyter Python Notebook.** Open a web browser by clicking on the web browser icon at the top of the toolbar:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/RCneZE7PEeaqTxIkdCEfsw_c491f272226b35805e44abef7a7a22a9_browser-icon.png?expiry=1515974400000&hmac=Z9UHKNyUUAZnhVoo8yrf-rw0t_YrEJK6TxIcyXGC3wo

Navigate to *localhost:8889/tree/Downloads/big-data-3*:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/TC6iOk7PEeaqTxIkdCEfsw_19f522873b2ae8800db929013dfa1913_notebook-browse.png?expiry=1515974400000&hmac=QFJfkOlhX82NVCWUzEkCLmvx-JW9sil7V7H_aFmwP6Q

Create a new Python Notebook by clicking on *New,* and then click on *Python 3:*



Step 2. **Load Pandas and Read a CSV file into a DataFrame.** We first load the Pandas library:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/DiAYqE7PEeaztg6Pg6w09w_382a062c2d05e19ec227ec529e3a4f89_import.png?expiry=1515974400000&hmac=NTPA2LruC7sJVcPFvaFTXnRTOfwe_vbjEL3k7eT1grE

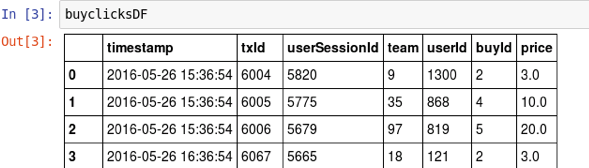
Note that to execute commands in Jupyter Notebooks, hold the *<shift>* key and press *<enter>.*

We can load the file *buy-clicks.csv* into a Pandas DataFrame:

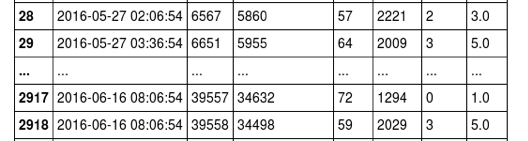
https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/A_FwW07PEea3kBK2xkPy7w_0355e2202cd657d25e079dd0e99745a3_readcsv.png?expiry=1515974400000&hmac=Xgpjo6OsTtT0REaShy98hoS-LLylescjLRC25R14-IE

This command assigns the DataFrame to a new variable named *buyclicksDF,* and reads the CSV using *pandas.read\_csv().*

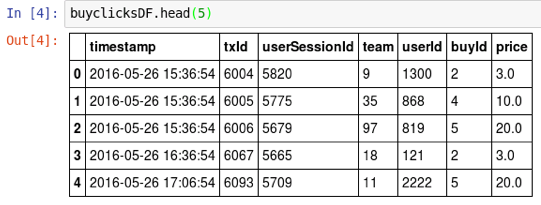
Step 3. **View the contents and shape of a DataFrame.** We can view the contents of the DataFrame by executing the variable:



Note that the Notebook does not display all the rows and displays the missing ones as *...*. :



We can view the first five rows by using the *head(5)* command:

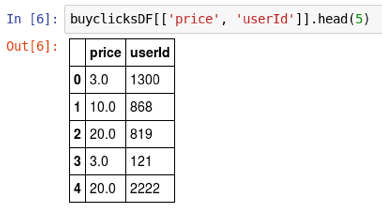


We can see how many rows and columns are in the DataFrame by looking at its shape:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/2BVoqE7OEeaubA6-qtnryw_ac4ed73f9844e5fd4ae3542b81d11df5_shape.png?expiry=1515974400000&hmac=ej_jWyfQqhPQzWdP5hbkHkzIQJXXEJl1sy2HqfilQSw

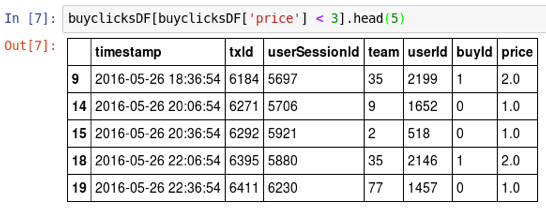
The result says that there are 2947 rows and 7 columns.

Step 4. **Filter rows and columns of a DataFrame.** We can view only the *price* and *userId* columns of the DataFrame:



The *[[ ]]* creates a copy of the DataFrame with only the specified columns.

We can also filter rows based on a criteria. The following selects rows with a price less than 3:



Step 5. **Calculate sum and average of a column.** Pandas DataFrames provide many aggregation operations. We can calculate the total price:



We can also calculate the average price:

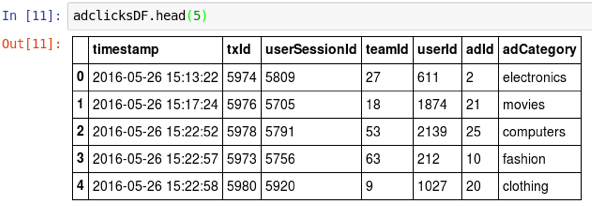
https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/cbbe307REea3kBK2xkPy7w_a671ee6f01cda24dfb7590c5591de3d7_mean.png?expiry=1515974400000&hmac=WCnbTRPLFlwQHN2gGTigC6OXwXRV5Pac7RQRkQ8I47w

A complete list of statistical aggregation operations for Pandas DataFrames is at <http://pandas.pydata.org/pandas-docs/stable/api.html#computations-descriptive-stats>

Step 6. **Combine two DataFrames.** We can combine two DataFrames on a single column. First, we will load *ad-clicks.csv* into a new DataFrame:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/33y3_k7SEeaughJrsEVARw_0a8944e993975c41c91bcd9266640ed0_read-adclicks.png?expiry=1515974400000&hmac=OKx0np24BOJqDLJmCSL8QR9bQUnKBp4hO6O0iFXL3BU

If we look at the contents, we see that *adclicksDF* also has a column named *userId:*



We can create a combine *buyclicksDF* and *adclicksDF* on the *userId* column with the following command:

https://d3c33hcgiwev3.cloudfront.net/imageAssetProxy.v1/6UIXG07SEeaX4QpLJOK7gQ_a49a6f98fb18fe62a2a5c37bec71b76c_merge.png?expiry=1515974400000&hmac=Xd5rutFJ-ewUUTgX7b-MhEOrnbukgvKQKlU9nlY3lM8

The combined DataFrame is assigned to a new variable named *mergeDF.* The command *adclicks.merge()* combines *adclicksDF* with the first argument *buyclicksDF*, and *on='userId'* denotes which column to join on.

We can see that the combined DataFrame contains the columns from both *adclicksDF* and *buyclicksDF:*

