10 HW #4B: Aggregation in Pandas

Repeat HW #4A, this time using Pandas. In order to receive full credit, please turn in a document which is python code containing what would be run to return the data asked.

Answer the following questions using only the syntax discussed in class. If a year is unspecified, please use the 2010 data and refer to the data dictionary for questions regarding the contents of the data.

Here are the statements that will load the data, note that you will need to change the directory.

```
## Initial Information
import pandas as pd
import numpy as np
df2010 = pd.read_csv('/Users/ncross/git/sqlnotes/newserver/data/2010.tdf',
        sep='\t', engine='python', names=['symb', 'retdate', 'opn', 'high', 'low',
        'cls' , 'vol', 'exch'])

df2011 = pd.read_csv('/Users/ncross/git/sqlnotes/newserver/data/2011.tdf',
        sep='\t', engine='python', names=['symb', 'retdate',  'opn', 'high', 'low',
        'cls', 'vol', 'exch'])

dffnd = pd.read_csv('/Users/ncross/git/sqlnotes/newserver/data/fnd.tdf',
        sep='\t', engine='python', names=['gvkey', 'datadate', 'fyear', 'indfmr',
        'consol', 'popsrc', 'datafmt' , 'tic', 'cusip', 'conm', 'fyr', 'cash', 'dp',
        'ebitda', 'emp', 'invt', 'netinc', 'ppent', 'rev' , 'ui', 'cik'])
```

The best approach to learning from these problems is to complete them using pen and paper, working by yourself and then using your group to double check your results. The First Five problems provide a short overview of the core concepts in the assignment, so make sure that you understand them. The Main Problems section contains questions which range from easy to very difficult. Remember to don't get stuck! If a problem is taking a long time or is too difficult, *use your group!*

First Five

1. What is the total number of rows in the 2010 database? Return this as an integer.
2. How many unique symbols are there in 2010? Return this as an integer.
3. What is the minimum closing price for a stock when it had a volume greater than 1,000,000 shares in 2010?
4. Return the average closing price for all stocks from the NYSE in 2010?
5. Return the average closing price for all stocks and the total number of rows by the exchange of each stock for 2010. Order the results from lowest to highest average closing price. This should return two rows and three index/value columns (exchange, average closing price and total number of rows).

Main Problems

1. Which symbols have less than 50 rows in 2010?
2. How many symbols have less than 50 rows in 2010?
3. Write a query which returns one row and two columns (DataFrame or Series). The first column should contain the number of symbols which have less than 50 rows in 2010 and the second column should have the number of symbols with more than 100 rows in 2010.

4. Write a query which returns two column and two rows (either series or a DataFrame). The first column should be equal to "lessThan50" or "moreThan100" and the second column should have the number of unique symbols which correspond to this condition. In other words, the same numbers as the previous problem, transposed with an column providing a description.

5. Write a query which returns three rows and two columns (note that one column maybe an index). One column should contain the average yearly total traded volume for symbols which had (1) more than 100 trading days (2) less than 50 trading days and (3) between 50 and 100 trading days. The other column should identify each row and be called "numType."

6. Write a query which returns three rows and two columns. The first column should contain the average *daily* traded volume for symbols which had (1) more than 100 trading days (2) less than 50 trading days and (3) between 50 and 100 trading days. The other column should identify each row and be called "numType."

7. How many of the symbols had a day where the dollar volume (closing price multiplied by number of shares traded) was greater than 100 million dollars in 2010?

8. What percentage of the symbols had a day where the dollar volume of shares traded was greater than 100 million dollars in 2010?