14 HW #6B: Pandas Joins (I)

Repeat HW #6A, 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.

The queries below rely on information from the stock return data. To load the data use the following commands. **Note: these load retdate as a date, rather than a string**

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!*

```
## Initial Information
import pandas as pd
import numpy as np

df2010D = pd.read_csv('/Users/ncross/git/sqlnotes/newserver/data/2010.tdf',
        sep='\t', engine='python', names=['symb', 'retdate', 'opn', 'high', 'low',
        'cls', 'vol', 'exch'], parse_dates=['retdate'])

df2011D = pd.read_csv('/Users/ncross/git/sqlnotes/newserver/data/2011.tdf',
        sep='\t', engine='python', names=['symb', 'retdate', 'opn', 'high', 'low',
        'cls', 'vol', 'exch'], parse_dates=['retdate'])

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'])

dfMTA = pd.read_csv('../sql-data/raw_data/mta/MTA_Hourly.tdf', sep='\t',
        engine='python', names=['plaza', 'mtadt', 'hr', 'direction', 'vehiclesez',
        'vehiclescash'])

dfTrans = pd.read_csv('../sql-data/raw_data/soapData.tdf', sep='\t',
        engine='python', names = ['orderid', 'userid', 'trans', 'type', 'local',
        'trans_dt', 'units', 'coupon', 'months', 'amt' ])
```

First Five

1. Using a join, create a dataset which contains symbol, the max closing price for that symbol from 2010 and the max closing price for that symbol from 2011. This should only include those symbol which are in both 2010 and 2011. Are you sure that both sides are unique? Why?

2. Using a LEFT JOIN, create a dataset which contains the following information: symbol, the last day it is traded in 2011 and the last day it is traded in 2010. Make sure to include all rows from 2011 and only those matching from 2010. There should be one row per symbol.

3. Using a cross join, create a dataset which contains every possible combination of symbol (in 2010) and return date (in 2010).

4. Write a query which returns the number of rows in the above query. How does this compare to the number of rows in the 2010 dataset? Does this make sense?

5. Write a query which has 12 rows and 3 columns. The first column should be Month (1,2,3...,12) the

second column should be the number of rows from that month in 2010 and the third column should be the number of rows from that month in 2011.

Main Problems

1. Using a LEFT JOIN, count the number of symbols which are in 2010, but not in 2011.
2. For each symbol, return the closing price on the first day that it is traded in 2010.
3. For each symbol, return the closing price on both the first day and last day that it is traded in 2010.
4. Create a dataset which contains 4 columns: the symbol, the retdate, the closing price and the closing price on the day after. Note that this dataset should *only* include Monday to Tuesday transitions, so retdate there should only be one row per-symbol per-Monday in the dataset. Specifically, if there are 50 trading weeks in a year and assuming that a symbol is traded every day, there would be 50 observations for that symbol
5. By matching the fnd data and the stocks 2010 data create a table which contains three columns and one row. The columns should represent the number of *unique* symbols which (a) are in both datasets, (b) are only in the 2010 dataset and (c) are only in the fnd data. Make sure to ignore all observations which are missing ticker symbols.
6. By combining the fnd and the stocks 2010 data, generate a dataset which contains the number of unique symbols of each of the three types in the previous problem. This time return two column and three rows (one of the columns should describe what data is in the row).
7. Create a dataset which is 5 rows by 3 columns. The first column should be DOW, the second column should be the average closing price of all stocks from 2010 on that day of the week and the third should be the average price of all stocks from 2011 for that day of the week.
8. We want to divide all stocks by the following criteria: if their max closing price in 2010 was less than 50, between 50 and 100 (inclusive) and more than 100. Return a table which contains the average net income (from fyear 2010) for each type of stock. Note that net income can be found in the fnd table and, if there are two net-income values for a particular ticker symbol, take the max. Only include those symbols in both datasets (fnd and s2010) and that do not have a missing net income.

Extra Problems

1. Create a dataset which contains the first day that each symbol is traded in 2010, the last day that the symbol is traded in 2011 and only includes those symbols which are in both 2010 and 2011.
2. For those symbols which had a closing price larger than \$100 *anytime* in 2010, return the symbol, the first day that it was traded in 2010 and all the dates that it had a closing price larger than \$200 in 2010. If the symbol was never above \$200, return no rows for it.
3. What are the first and last date listed for each symbol in 2010? Be careful to return this for *each* symb.
4. For each symbol that appears anywhere in 2010, calculate the number of missing trading days that it has in each month in 2010. This should return three columns: symbol, month, number of missing values.
5. Create a dataset which is 10 rows by 3 columns. The first column should be the year, the second column should be the day-of-the-week and the third column should be the average closing price of all stocks for that day-of-the-week. Include both 2010 and 2011.