13 HW #6A: SQL Joins (I)

The following questions utilize the financial data in the s2010, s2011 and fnd tables. Before beginning the assignment, *please read the data dictionary to better understand the data.* When doing so, keep an eye on data types for different columns as well as table organization.

- If no table information is given, use the 2010 data.
- If the query returns a significant number of rows, please only copy a few rows in your response.

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. 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 symbols 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 columns 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) 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, 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.

6. How many cars (total), on an average *day*, go through each toll plaza in both directions combined (return a row for each toll plaza)? Make sure to sum up to the *day* level before computing the average.

7. Which day-of-the-week (Monday, Tuesday, etc.) has the highest number of cars going through Plaza #1, both directions combined, with EZ pass? This should be the total number of cars over the entire time period in the dataset.

8. Which day-of-the-week-plaza combination has the lowest percentage of users cars using the EZ pass in the outbound direction? In other words, if you look at outbound cars through each plaza, which day of the week has the lowest percentage of cars using EZ pass. You can compute the percentage over the entire time period.

9. Calculate the average number of cars going through Plaza #1, outbound, with EZ pass for each day-of-the-week. This should be a *daily* average and should return 7 rows.

10. In an average week on Plaza #1 with EZ pass (outbound), what percentage of cars go through each day? (E.g. basically the above, but this time percent of total).

11. For each plaza, what was the change (percent) in average number of cars on a Monday using EZ-pass in both directions, between 2015 and 2016? (Calculate the average number of cars for a Monday in 2015 and 2016 and then calculate the percentage change based off of that.)

12. Calculate, for each hour, plaza and day-of-the-week (so $7 \cdot 24$ rows per plaza), the ratio of inbound to outbound traffic.

13. Using a join, create a dataset with three columns and 7 rows. The first column should be the DOW, the second column should be the average number of cars, per-day-of-the-week, through toll Plaza #1 in either direction with an EZ pass in 2016 and the final column should be the average number of cars, per-day-of-the-week, through tool Plaza #2 with an EZ pass in 2015.

14. Create a dataset which contains twenty-four rows and two columns. The first column represents the hour and the second column represents the max number of EZ pass cars, during that hour, outbound, through Plaza #1.

15. Create a dataset which contains 24x7 rows and two columns. The first column represents the DOW-hour combination (you may need to combine two columns using "||" or the concatenate operator) and the second represents the max number of EZ pass cars, during that hour-day, through Plaza #1 in the outbound direction.

16. Using at least one join, create a dataset which contains twenty-four rows and 4 columns. Each row should represent an hour, and the first column should be an hour identifier. Column #2 should contain the maximum number of EZ pass cars, in the inbound direction, through Plaza #1 during that hour, Column #3 should contain the minimum number of outbound EZ-pass cars, during that hour, through Plaza #2 and Column #3 should be the maximum number of EZ-pass cars in either directions combined, during that hour, on Plaza 3.

17. Create a dataset which contains the following columns: hour, day-of-the-week, plaza, the ratio of inbound to outbound traffic in 2014 and the ratio of inbound to outbound traffic in 2013.

18. For the day with the most traffic (inbound, outbound and both payment types combined), calculate the ratio of inbound to outbound traffic over the entire dataset (not by plaza), for each hour. Return three columns, the day-of-the-week of that date, hour and the percent for that hour.