6 HW #4A: Aggregate Functions and Dates

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.

1. Using the daily stock return data answer the following questions:
   (a) What is the total number of rows in the 2010 database?
   (b) How many unique cusips are there in 2010?
   (c) How many unique cusips are there in 2011?
   (d) Which cusips have less than 50 rows in 2010?
   (e) How many cusips have less than 50 rows in 2010?
   (f) Write a query which returns one row and two columns. The first column should contain the number of cusips which have less than 50 rows in 2010 and the second column should have the number of cusips with more than 100 rows in 2010.
   (g) Write a query which returns two column and two rows. The first column should be named "numtype" which should be equal to "less than 50" or "more than 100" and the second column should have the number of unique cusips which correspond to this condition. In other words, the same numbers as the previous problem, transposed with an column providing a description.
   (h) Write a query which returns three rows and two columns. The first column should contain the average yearly total traded volume for cusips 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."
   (i) Write a query which returns three rows and two columns. The first column should contain the average *daily* traded volume for cusips 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."
   (j) How many of the permnos had a day where the dollar volume of shares traded was greater than 100 million dollars in 2010?
   (k) What percentage of the permnos had a day where the dollar volume of shares traded was greater than 100 million dollars in 2010?

HC

2. More Advanced Questions: Answer the following questions using the d2010 table:
   (a) Which day of the week (0,1,2,...) had the largest number of shares traded?
   (b) Which day of the week (0,1,2,..) has the highest average shares traded?
   (c) Which day of the week-month (January-Monday, January-Tuesday, etc.) combination had the highest average return? Note that both day of the week and month can be kept as integers.

SS

   (d) Write a query which returns 3 columns and 5 rows with each row should represent a day of the week. One column should be the English day of the week ("Monday,", "Tuesday," etc.) while the next column should be equal to the average number of shares traded on that day from stocks that have a volume traded between 1 million and 2 million shares on that day ("C2"). The final column ("C3") should be the average number of shares traded on that day from stocks that had a volume traded outside of 1 million to 2 million.

7 HW #4B: Aggregation and Dates

The queries below rely on information from both the stock return data and the California traffic data (stocks2016 and cls.traffic).

1. The following questions deal with the stock data.
   (a) Which quarter in 2010 has the most trading days?[7]
   (b) Write a query which returns the maximum price for each permno in 2010, making sure to ignore NULL prices as well as sorting by price from high-to-low.
   (c) Write a query which returns permno and a column "DFlag", which is equal to 1 if the max price (ignoring Nulls) in 2010 is larger than 100, 2 if the max price in 2010 is between 50 and 100 and 3 if the max price is less than 50.
   (d) Write a query which returns the number of distinct permnos of each type of Dflag. This should be 3 rows and 2 columns (one of the columns should indicate what each row means).
   (e) Write a query which returns the number of distinct permnos of each type of Dflag, this should be 3 columns and a single row.
   (f) Calculate the number of distinct trading days per month in 2010. This should return 12 rows with 2 columns.
   (g) For each permno, calculate the difference between the maximum and minimum price for December, 2010 once again removing Null prices. Only include those stocks with 22 observations (there are 22 trading days in December, 2010).
   (h) Calculate the average difference between the maximum and minimum price for Tuesdays in January, 2010, once again removing Null prices. Only include those stocks with 4 observations which fulfill the criteria.[8]
   (i) Same as before, but for Wednesday. Calculate the average price for Wednesday in January 2010, once again removing Null prices. Only include those stocks with 4 observations which fulfill the criteria.
   (j) Calculate the average price for Tuesday in January 2010, once again removing Null prices. Only include those stocks with 4 observations which fulfill the criteria.
   (k) For those stocks which have four Tuesday, January 2010 observations, calculate the number which have an average price larger than the average of all Wednesday, January 2010 prices and the number which have less than the average of all Wednesday prices. When calculating the average Wednesday only include those stocks with four wednesday observations after removing Null prices.

KK

2. The following questions deal with the California traffic data.
   (a) Using a WHERE clause to filter out rows, return the average swpeakhr for counties that begin with the letter "A"
   (b) Using a WHERE clause to filter out rows, return the average swpeakhr for counties that begin with the letter "A", this time breaking it down by county.
   (c) Return the average swpeakhr for counties that begin with the letter "A" or the letter "M".

JL

[7] Define Q1 as Jan-Mar, Q2 as Apr-Jun, etc.
[8] There are 4 Tuesday trading days in January, 2010.

(d) Return two columns and a single row: One which contains the average swpeakhr for those counties which begin with "A" and one which contains the average swpeakhr for those counties that being with "M".

(e) Return two columns: One which contains the average swpeakhr for those counties which begin with "A" and one which contains the average swpeakhr for those counties that being with "M". The rows should be grouped by county and there should be a row for each county that begins with either "A" or "M".

(f) Write a query which returns the number of distinct ROUTES with more than 5,000 swpeakhr or less than 2,000 nepeakhr for each county:

JL

(g) What is average ratio of nepeakhr to swpeakhr (by county and routeno)?

(h) What county has the highest percentage of routes with an average nepeakhr to swpeakhr ratios less than 1?

(i) Of the county-route combinations which have average nepearkhr to swpeakhr ratio less than 1, return the breakdown of if their average swavgday is less than 5,000 or more than 5,000.

(j) Of the counties with a swpeakhr of more than 15000, provide a breakdown of the max swpeakhr for each county-route combination (is it more than 10K, 5K or 0?).

8 HW #4C: Information Schema and Exploring the Relationship between Returns and Dollar Volume

In the following exercise, we will investigate the relationship between the dollar volume of shares traded and the returns of that company.

SS

1. Using the information schema answer the following questions:
   (a) Write a query which returns the count of data types (int, float, etc.) of each columns in the stock2016 schema.
   (b) Write a query which returns the number of distinct column types in the entire database.
   (c) Write a query which returns 3 columns: schema name, column data type, and the number of columns in that schema of that column type.
   (d) Rewrite the above query in a wide-format. Each row should represent a single schema.
   (e) Create a pie chart of the above information for the schema "information_schema," which data format (wide or long) did you use?

SM

2. Exploring the relationship between dollar volume and return:
   (a) Write a query which returns the return rounded to the nearly thousandth of a percent while dealing with any data issues. Return the data in hundredths, so if the return is .037123, 3.7 should be returned. Include the dollar volume of stocks traded that day, rounded to the nearest 1,000. Also, only take a 1/16 sample using the following where statement:

   ```
   where md5( permno::varchar(100) ) like '0%'
   ```

   (b) Create a scatter plot of your rounded returns vs. the rounded volume.
   (c) Run simple linear regression on the rounded returns vs. the rounded volume and report the results. Do you believe that there is a relationship between trading size and volume traded?
   (d) Recreate the scatter plot making sure to remove days with less than 250 million shares traded and only include returns between -10 and 10. Did the pattern change?
   (e) Run simple linear regression on the rounded returns vs. the rounded volume and report the results for the sample of less than 250 million shares and returns between -10 and 10. Do you believe that there is a relationship between trading size and volume traded?
   (f) Using only the SUM, AVG and COUNT aggregate functions, compute the variance of both the rounded volume and the rounded returns of the sample.
   (g) Using only the SUM, AVG and COUNT aggregate functions, compute the covariance between the the rounded returns and volume.