3 HW #3: Aggregate Functions

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 Stocks 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 observations in 2010?
   (e) How many cusips have less than 50 observations 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 observations in 2010 and the second column should have the number of cusips with more than 100 observations 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?

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

   (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) 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?

   (d) Using only the SUM, AVG and COUNT aggregate functions, compute the variance of both the rounded volume and the rounded returns of the sample.

   (e) Using only the SUM, AVG and COUNT aggregate functions, compute the covariance between the the rounded returns and volume.

3. 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?

4. Answer the following questions using the 2010 data:

   (a) Which day of the week (Monday, Tuesday, etc.) had the largest number of shares traded?

   (b) Which day of the week (Monday, Tuesday, etc.) has the highest average shares traded?

(c) Which day of the week-month (January-Monday, January-Tuesday, etc.) combination had the highest average return?

(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. The final column 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.