SQL Stretch

Additional Content

https://classroom.udacity.com/courses/ud198 is a free (though you will need to create an account) online course on SQL including sections on topics we don't cover, namely window functions and performance tuning. If these are topics you don't know about, then you might want to look through this material.

Technical exercises

Nested Subqueries

Your challenge is to investigate the amount of monthly variation in sales at each store within the Iowa liquor database. More specifically, for each store you need to find the difference between largest monthly sales and lowest monthly sales. You then need to find how big this difference was on average and as max and min across all stores.

Correlation

Your challenge is to calculate the (Pearson) correlation between the "pack" and "btl_price" fields in the sales table of the Iowa liquor database *without using the built-in CORR() function.* You can of course use CORR() to check your answer.

As a bonus, also calculate the Spearman's rank correlation between these two variables. For this you can use CORR(), but you'll have to adapt it.

You may not know what these correlation coefficients are or how to calculate them. In this case, read on (and check Google!):

If you measure two features (*x* and *y*) for a number of data points (e.g., the height and weight of a number of people), the correlation between those two variables describes how closely an increase in one corresponds to an increase in the other.

The most common way of measuring correlation is Pearson's correlation coefficient, usually given the symbol *r* (and if someone says correlation without specifying, you can assume they mean Pearson's coefficient). Pearson's coefficient is defined by the formula

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

In words, you calculate the mean of the *x* values and subtract this from every *x* value to get the "*x* deviations" (how far each *x* value deviates from the mean) and do the same for the *y* values to find the *y* deviations. Then you multiply each *x* deviation by the corresponding *y* deviation and add them all up to get the numerator. To get the denominator, you square each *x* deviation and all them all up, square each *y* deviation and add them all up, then multiply together these two sums of squared deviations, and square root. Finally, you divide the numerator by the denominator to find *r*.

So you need to implement this calculation in SQL. There are probably many ways to do this, so I'd be interested to see what different ways people come up with.

Pearson's coefficient measures *linear* relationships between two variables: relationships of the form "if *x* increases by a fixed amount, *y* increases (or decreases) by a fixed amount." Sometimes, variables have a non-linear relationship, but it is still true that an increase in *x* corresponds to an increase in *y*.

Spearman's rank correlation coefficient is an alternative way of measuring correlation that allows for non-linear relationships. Spearman's correlation is the Pearson correlation between the *ranks* of the variables, rather than the variables themselves. The ranks are computed by putting the variables in order and setting the smallest to have rank 1, the next to have rank 2, and so on. So Spearman's coefficient only cares what order the values come in, without looking at how big the gaps between them are.

Chinook Database

At the link below, you will find instructions to access a SQL database with information about a fictional music store, and a selection of practice questions (not sure about the difficulty).
https://docs.google.com/document/d/1mCLJCdGcW5n4RSjCvF8mXyQzGRD6eTWuaRkzCTDZ9FE/

Hackerrank

At https://www.hackerrank.com/domains/sql, (need to sign up, but it's free) you will find a range of SQL practise problems, rated as easy, medium, or hard (only 1 hard, but some of the medium ones look pretty challenging).

Data Interpretation

The following questions are more open-ended questions about using SQL to explore the data and derive insight. They are vague "real-world" questions, and your job is to figure out how to translate them into queries you can apply to the database and interpret the results. They all apply to the Iowa liquor database.

For some of these, there may not be a single query that unambiguously answers the question, so you should consider several queries and see if they all point to the same answer, or if they give different answers, in which case you should critically compare them and decide on the most useful.

Some of these questions may require you to find additional data from other sources; how deep you go into doing that is up to you.

1. Which store is the most profitable (assuming all stores have the same overheads)?
2. Which vendor is most profitable (same assumption)?
3. If the state of Iowa were to run a campaign aimed at reducing people's alcohol intake, which counties should they target?
4. At what time of year should alcohol stores run ad campaigns? Where should they run them?
5. Which are more popular: American or imported drinks?
6. What factors lead to a store being inactive?
7. Do drinking habits vary between urban and rural areas?
8. How do sales vary over the course of a quarter? A month? A week? What is the most useful time period over which to study sales?