Practice problem set: Joins and Aggregation

This problem set contains sixteen practice problems that exercise joins and aggregation. These problems take the place of a reading for this week. You will *not* be asked to hand in your answers for credit, but I or the TA will be happy to check your answers and give you feedback if you ask. NOTE: problems like these make ideal midterm questions.

Consider the following schema:

Region(regionkey, rname)
Nation(nationkey, regionkey, nname)
Supplier(suppkey, snationkey, sname)
Part (partkey, ptype)
PartSupp(partkey, suppkey, unit_cost)
Customer(custkey, cnationkey, cname, m_segment)
Orders(orderkey, custkey, orderdate, priority)
LineItem(orderkey, line_num, partkey, suppkey, qty, ext_price, commit_date, ship_date, receipt_date, return_flag)

Create RA expressions that answer the following questions:

Q1: Find the average and total value of items shipped from Estonia during 2012. Exclude items that were returned.

LineItem \(\sigma\) year(ship_date)=2012 and not return_flag
\(\Gamma\) sum(ext_price), avg(ext_price)

Q2: Find, for a part of type “Widget”, the supplier in region “Asia” who offers the lowest unit cost. Do not worry about ties.

Part \(\sigma\) ptype='Widget' \(\join\) PartSupp on partkey
\(\join\) Supplier on suppkey \(\join\) Nation on nationkey=snationkey
\(\join\) Region on regionkey \(\sigma\) rname='Asia' \(\tau\) unit_cost \(\lim\) 1

Q3: Who are the “big fish” customers? List every customer that has ever placed an order whose combined line item quantities totaled 1000 or higher.
Q4: Find, for market segment “manufacturing”, the ten orders that represented the most potential revenue at year end 2012, where potential revenue of an order is the value of its unshipped items. Do not worry about ties.

Q5: Unsold inventory is expensive. The supplier had to pay for it with money that could have been invested better somewhere else (lost profits), and the item takes up space that might be put to better use (more lost profits). No grocery store would ever dedicate half its floor space to ramen noodles, for example. There’s just not enough potential profit in cheap soup to justify buying that amount of inventory, or dedicating that amount of floor space to it. For supplier “Acme”, identify any/all parts where the quantity in stock is more than half a year’s supply, for the year 2013 (give the part, the quantity on hand, and the quantity sold last year). In other words, if 100 widgets were sold last year, and there are currently 50 or more in stock, we want to know about it.

Q6: Suppose we are looking for ways to improve customer satisfaction. For each order priority, what fraction of orders, made in 2011, contained at least one item which arrived later than the commit date when receipt was promised?

Q7: Suppose we want to partner with new local suppliers to reduce our international shipping costs. For the year 2010, and for all nations in region “Americas”, what fraction of supplier revenue (value of orders shipped) was generated by customers in the same nation as the supplier?
Q8: For each of the years 2009 through 2013, and for parts of type “Widget”, what was the market share of Canada? (Market share is the fraction of total sales due to suppliers in that nation).

Q9: What nations sustain the business? Give a breakdown of total profits by nation, most profit first, for the year 2012. Profit is defined for a single line item as the extended price minus the product of supplier’s unit cost and line item quantity.

Q10: If a customer returns an item they ordered, the corresponding revenue is lost. Which customers caused the most lost revenue in 2012?

Q11: Given that the human effort to process an order doesn’t depend very strongly on the number of items ordered, we generally prefer to work with larger orders (it’s actually easier to ship a whole case unopened than to open a case and pull out one item, for example). However, small orders are also more common than large ones. What fraction of 2012 revenue was due to “small” orders with fewer than ten total items shipped? Put another way, if we were a grocery store, what fraction of revenue came from the “express checkout” line?

Q12: Which supplier generated the most revenue in 2012?
Q13: It would be very bad for business to lose any single supplier that was responsible for at least 10% of yearly revenue in 2012. Which suppliers are Too Big to Fail?

Q14: Suppose our contract with a shipping partner is about to expire and we need to re-negotiate shipping rates for merchandise traveling between Portugal and Brazil. What is the total value of goods that traveled between the two nations, in either direction, for each of the years 2010 through 2013, inclusive?

Q15: An inventory valuation determines what items sitting in a supplier’s warehouses represent a significant fraction of the supplier’s total assets. For suppliers in India, what in-stock parts represent at least 2% of total inventory value?

Q16: In order to target marketing better, it helps to know how frequently our customers order from us. A binary histogram can be very helpful: it shows how many customers made a number of orders rounded down to the nearest power of two (e.g. 0, 1, 2-3, 4-7, 8-15, ...). Suppose we have available the function log2i(x), which returns the largest k = 2**i where k <= x (or 0 for x=0). Use log2i() to bucket customers in a binary histogram by how many orders they made in 2011. For example, 10 customers might land in bucket 0 (no orders), 20 customers might land in bucket 4 (4-7 orders), etc.