Practical Exam: Grocery Store Sales

FoodYum is a grocery store chain that is based in the United States.

Food Yum sells items such as produce, meat, dairy, baked goods, snacks, and other household food staples.

As food costs rise, FoodYum wants to make sure it keeps stocking products in all categories that cover a range of prices to ensure they have stock for a broad range of customers.

Data

The data is available in the table `products`.

The dataset contains records of customers for their last full year of the loyalty program.

Column Name	Criteria
product_id	Nominal. The unique identifier of the product. Missing values are not possible due to the database structure.
product_type	Nominal. The product category type of the product, one of 5 values (Produce, Meat, Dairy, Bakery, Snacks). Missing values should be replaced with “Unknown”.
brand	Nominal. The brand of the product. One of 7 possible values. Missing values should be replaced with “Unknown”.
weight	Continuous. The weight of the product in grams. This can be any positive value, rounded to 2 decimal places. Missing values should be replaced with the overall median weight.
price	Continuous. The price the product is sold at, in US dollars. This can be any positive value, rounded to 2 decimal places. Missing values should be replaced with the overall median price.
average_units_sold	Discrete. The average number of units sold each month. This can be any positive integer value. Missing values should be replaced with 0.
year_added	Nominal. The year the product was first added to FoodYum stock. Missing values should be replaced with 2022.
stock_location	Nominal. The location that stock originates. This can be one of four warehouse locations, A, B, C or D Missing values should be replaced with “Unknown”.

Task 1

Last year (2022) there was a bug in the product system. For some products that were added in that year, the `year_added` value was not set in the data. As the year the product was added may have an impact on the price of the product, this is important information to have.

Write a query to determine how many products have the `year_added` value missing. Your output should be a single column, `missing_year`, with a single row giving the number of missing values.

index	missing_year
0	

Rows: 1

Task 2

Given what you know about the year added data, you need to make sure all of the data is clean before you start your analysis. The table below shows what the data should look like.

Write a query to ensure the product data matches the description provided. Do not update the original table.

Column Name	Criteria
product_id	Nominal. The unique identifier of the product. Missing values are not possible due to the database structure.
product_type	Nominal. The product category type of the product, one of 5 values (Produce, Meat, Dairy, Bakery, Snacks). Missing values should be replaced with “Unknown”.
brand	Nominal. The brand of the product. One of 7 possible values. Missing values should be replaced with “Unknown”.
weight	Continuous. The weight of the product in grams. This can be any positive value, rounded to 2 decimal places. Missing values should be replaced with the overall median weight.

Column Name	Criteria
price	Continuous. The price the product is sold at, in US dollars. This can be any positive value, rounded to 2 decimal places. Missing values should be replaced with the overall median price.
average_units_sold	Discrete. The average number of units sold each month. This can be any positive integer value. Missing values should be replaced with 0.
year_added	Nominal. The year the product was first added to FoodYum stock. Missing values should be replaced with last year (2022).
stock_location	Nominal. The location that stock originates. This can be one of four warehouse locations, A, B, C or D Missing values should be replaced with “Unknown”.

i	prod…	product_…	brand	\	|	average_units_sold	year…	stock_location
32	33	Snacks	YumMie	445.79	6.49	18	2017	D
36	37	Snacks	YumMie	299.46	5.81	18	2018	B
53	54	Bakery	YumMie	559.66	7.01	16	2015	C
80	81	Bakery	YumMie	401.34	6.85	16	2019	A
98	99	Snacks	YumMie	423.91	6.4	18	2017	B
104	105	Bakery	YumMie	487.64	6.87	16	2020	A
155	156	Bakery	YumMie	454.11	7.19	16	2020	D
168	169	Snacks	YumMie	419	6.11	17	2019	A
199	200	Bakery	YumMie	614.85	7.36	16	2016	C
213	214	Snacks	YumMie	408.17	6.23	18	2019	B
220	221	Bakery	YumMie	427.91	7.1	16	2019	A
228	229	Meat	YumMie	420.59	12.04	25	2021	B
268	269	Produce	YumMie	538.12	4.08	21	2021	C
292	293	Dairy	YumMie	697.9	9.07	22	2021	B
298	299	Snacks	YumMie	477.15	5.86	18	2022	C
307	308	Dairy	YumMie	577.78	8.95	22	2022	B

Rows: 1,700

Task 3

To find out how the range varies for each product type, your manager has asked you to determine the minimum and maximum values for each product type.

Write a query to return the `product_type`, `min_price` and `max_price` columns.

index	product_type	min_price	max_price
0	Snacks	5.2	
1	Produce	3.46	
2	Dairy	8.33	
3	Bakery	6.26	
4	Meat	11.48	

Rows: 5

Task 4

The team want to look in more detail at meat and dairy products where the average units sold was greater than ten.

Write a query to return the `product_id`, `price` and `average_units_sold` of the rows of interest to the team.

index	product_id	price	average_units_sold
0	6	16.2	
1	8	15.77	
2	9	11.57	
3	10	13.94	
4	11	9.26	
5	14	11.92	