PROJECT: PREPARE FOR DATA ANALYST ASSOCIATE CERTIFICATION

Practical Exam Sample: Pet Supplies

PetMind is a retailer of products for pets. They are based in the United States.

PetMind sells products that are a mix of luxury items and everyday items. Luxury items include toys. Everyday items include food.

The company wants to increase sales by selling more products for some animals repeatedly.

They have been testing this approach for the last year.

They now want a report on how repeat purchases impact sales.

Data

The data is available in the table `pet_supplies`.

The dataset contains the sales records in the stores last year.

Column Name	Criteria
product_id	Nominal. The unique identifier of the product. Missing values are not possible due to the database structure.
category	Nominal. The category of the product, one of 6 values (Housing, Food, Toys, Equipment, Medicine, Accessory). Missing values should be replaced with “Unknown”.
animal	Nominal. The type of animal the product is for. One of Dog, Cat, Fish, Bird. Missing values should be replaced with “Unknown”.
size	Ordinal. The size of animal the product is for. Small, Medium, Large. Missing values should be replaced with “Unknown”.
price	Continuous. The price the product is sold at. Can be any positive value, round to 2 decimal places. Missing values should be replaced with the overall median price.
sales	Continuous. The value of all sales of the product in the last year. This can be any positive value, rounded to 2 decimal places. Missing values should be replaced with the overall median sales.
rating	Discrete. Customer rating of the product from 1 to 10. Missing values should be replaced with 0.
repeat_purchase	Nominal. Whether customers repeatedly buy the product (1) or not (0). Missing values should be removed.

Unknown integration DataFrame as `Overview`

```
--- Overview of pet_supplies table

SELECT *
FROM pet_supplies
LIMIT 10;
```

i...	produc...	cate...	a..	s...	p..	s...	r...	repeat_purchase
0	1	Food	Bird	large	51.1	1860.62	7	
1	2	Housing	Bird	MEDIUM	35.98	963.6	6	
2	3	Food	Dog	medium	31.23	898.3	5	
3	4	Medicine	Cat	small	24.95	982.15	6	
4	5	Housing	Cat	Small	26.18	832.63	7	
5	6	Housing	Dog	Small	30.77	874.58	7	
6	7	Housing	Dog	Small	31.04	875.07	5	
7	8	Toys	Cat	medium	28.9	1074.31	4	
8	9	Equipment	Fish	MEDIUM	17.82	503.67	5	
9	10	Medicine	Dog	medium	24.93	838.88	8	

Rows: 10 Expand

Unknown integration DataFrame as `Category`

```
--- Exploring category column for NULLs, empty strings, missing, or inconsistent data

SELECT DISTINCT category
FROM pet_supplies;
```

index	category
0	Medicine
1	Food
2	Equipment
3	-
4	Accessory
5	Housing
6	Toys

Rows: 7 Expand

Unknown integration DataFrame as `A`

```
--- Exploring animal column for NULLs, empty strings, missing, or inconsistent data

SELECT DISTINCT animal
FROM pet_supplies;
```

0	Fish
1	Cat
2	Bird
3	Dog

Rows: 4 Expand

Unknown integration DataFrame as `S`

```
--- Exploring size column for NULLs, empty strings, missing, or inconsistent data

SELECT DISTINCT size
FROM pet_supplies;
```

0	large
1	medium
2	Large
3	Medium
4	SMALL
5	small
6	LARGE
7	MEDIUM
8	Small

Rows: 9 Expand

Unknown integration DataFrame as

```
--- Exploring price column for NULLs, empty strings, missing, or inconsistent data

SELECT price
FROM pet_supplies
WHERE price ILIKE '%Un%';
```

	price
0	unlisted
1	unlisted
2	unlisted
3	unlisted
4	unlisted
5	unlisted
6	unlisted
7	unlisted
8	unlisted
9	unlisted
10	unlisted
11	unlisted
12	unlisted
13	unlisted
14	unlisted
15	unlisted

Rows: 150 Expand

Unknown integration DataFrame as

```
--- Exploring sales column for NULLs, empty strings, missing, or inconsistent data

SELECT product_id, sales
FROM pet_supplies
WHERE sales IS NULL;
```

Your query ran successfully but returned no results.

Unknown integration DataFrame as

```
--- Looking for Max, Min, Avg and Sum sales in our dataset

SELECT
    MIN(sales),
    MAX(sales),
    AVG(sales),
    SUM(sales)
FROM pet_supplies;
```

	min	max	avg	s...
0	286.94	2255.96	996.5978466667	1494896.77

Rows: 1 Expand

Unknown integration DataFrame as

```
--- Exploring rating column for NULLs, empty strings, missing, or inconsistent data

SELECT DISTINCT rating
FROM pet_supplies;
```

0	8
1	9
2	
3	7
4	1
5	5
6	2
7	4
8	6
9	3

Rows: 10 Expand

Unknown integration DataFrame as

```
--- Exploring rating repeat_purchase for NULLs, empty strings, missing, or inconsistent data

SELECT DISTINCT repeat_purchase
FROM pet_supplies;
```

	repeat_pu…
0	0
1	1

Rows: 2 Expand

Unknown integration DataFrame as

```
--- Understanding the table structure and data types of all the columns

SELECT column_name, data_type
FROM information_schema.columns
WHERE table_name = 'pet_supplies';
```

	column_n…	data_type
0	product_id	integer
1	category	text
2	animal	text
3	size	text
4	price	text
5	sales	double precision
6	rating	integer
7	repeat_purchase	integer

Rows: 8 Expand

We have checked each columns for nulls, mistakes, incosistencies or missing values. Here are the problems we need to solve:

1. Product_id: PERFECT! just make sure the datatype.
2. Category: Certain products are categorised as '-' which means missing data.
3. Animal: PERFECT!
4. Size: Extra categories - capitalizations issues
5. Price: Certain products are tagged as 'unlisted' which means missing data.
6. Sales: PERFECT! incase any value is missing just give it default median value
7. Rating: Missing Values - No ratings are given for certain products
8. Repeate_purchase: PERFECT!

Make sure to CAST everything properly!

Task 1

From taking a quick look at the data, you are pretty certain it isn't quite as it should be. 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 return a table that 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.
category	Nominal. The category of the product, one of 6 values (Housing, Food, Toys, Equipment, Medicine, Accessory). Missing values should be replaced with “Unknown”.
animal	Nominal. The type of animal the product is for. One of Dog, Cat, Fish, Bird. Missing values should be replaced with “Unknown”.
size	Ordinal. The size of animal the product is for. Small, Medium, Large. Missing values should be replaced with “Unknown”.
price	Continuous. The price the product is sold at. Can be any positive value, round to 2 decimal places. Missing values should be replaced with 0.
sales	Continuous. The value of all sales of the product in the last year. This can be any positive value, rounded to 2 decimal places. Missing values should be replaced with the overall median sales.
rating	Discrete. Customer rating of the product from 1 to 10. Missing values should be replaced with 0.
repeat_purchase	Nominal. Whether customers repeatedly buy the product (1) or not (0). Missing values should be removed.

Unknown integration DataFrame as

```
--- CTE sales_median

WITH sales_median AS (
    SELECT
        PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY sales) :: NUMERIC AS median_sales
    FROM public.pet_supplies
    WHERE sales IS NOT NULL
)

-----------------------------------------------------------------------------------------------------------------------

SELECT
    product_id::INTEGER,

    -- Replace '-' and NULL in category with 'Unknown'
    COALESCE(NULLIF(category, '-'), 'Unknown') AS category,

    --- No Missing values
    animal,

    -- Normalize size capitalization and replace invalid or NULL with 'Unknown'
    CASE
        WHEN UPPER(size) IN ('SMALL', 'MEDIUM', 'LARGE') THEN INITCAP(LOWER(size))
        ELSE 'Unknown'
    END AS size,

    -- Replace 'unlisted' or NULL price with 0, cast to numeric, round 2 decimals
    ROUND(COALESCE(NULLIF(price, 'unlisted') :: NUMERIC, 0), 2) AS price,

    -- Replace NULL sales with median sales from CTE, round 2 decimals
    ROUND(COALESCE(sales, (SELECT median_sales FROM sales_median)) :: NUMERIC, 2) AS sales,

    --- Replace NULL values with 0
    COALESCE(rating, 0)::INTEGER AS rating,

    -- Filtered in WHERE clause, so no COALESCE here
    repeat_purchase

FROM pet_supplies
WHERE repeat_purchase IN (0, 1);
```

	p...	c...						repeat_pu...
0	1	Food	Bird	Large	51.1	1860.62	7	1
1	2	Housing	Bird	Medium	35.98	963.6	6	0
2	3	Food	Dog	Medium	31.23	898.3	5	1
3	4	Medicine	Cat	Small	24.95	982.15	6	1
4	5	Housing	Cat	Small	26.18	832.63	7	1
5	6	Housing	Dog	Small	30.77	874.58	7	0
6	7	Housing	Dog	Small	31.04	875.07	5	0
7	8	Toys	Cat	Medium	28.9	1074.31	4	0
8	9	Equipment	Fish	Medium	17.82	503.67	5	0
9	10	Medicine	Dog	Medium	24.93	838.88	8	0
10	11	Food	Dog	Large	40.87	1457.22	7	1
11	12	Medicine	Bird	Medium	34.96	1204.6	5	1
12	13	Food	Dog	Medium	31.07	889.73	4	0
13	14	Food	Dog	Large	40.8	1450.5	6	1
14	15	Accessory	Bird	Medium	33.13	859.29	4	1
15	16	Accessory	Bird	Large	43.09	1418.72	1	1

Rows: 1,500 Expand

Task 2

You want to show whether sales are higher for repeat purchases for different animals. You also want to give a range for the sales.

Write a query to return the `animal`, `repeat_purchase` indicator and the `avg_sales`, along with the `min_sales` and `max_sales`. All values should be rounded to whole numbers.

You should use the original `pet_supplies` data for this task.

Unknown integration DataFrame as

```
--- animal_sales

SELECT
    animal,
    repeat_purchase,
    ROUND(AVG(sales) :: NUMERIC) AS avg_sales,
    ROUND(MIN(sales) :: NUMERIC) AS min_sales,
    ROUND(MAX(sales) :: NUMERIC) AS max_sales

FROM pet_supplies
GROUP BY animal, repeat_purchase
ORDER BY animal, repeat_purchase;
```

		repeat_pu...	a	m	m
0	Bird	0	1380	858	2255
1	Bird	1	1408	853	2256
2	Cat	0	1035	512	1730
3	Cat	1	998	512	1724
4	Dog	0	1084	574	1795
5	Dog	1	1038	574	1797
6	Fish	0	705	288	1307
7	Fish	1	693	287	1301

Rows: 8 Expand

Task 3

The management team want to focus on efforts in the next year on the most popular pets - cats and dogs - for products that are bought repeatedly.

Write a query to return the `product_id`, `sales` and `rating` for the relevant products.

You should use the original `pet_supplies` data for this task.

Unknown integration DataFrame as

```
--- popular_pet_products

SELECT
    product_id,
    sales,
    rating

FROM pet_supplies
WHERE (animal = 'Cat' OR animal = 'Dog')
    AND repeat_purchase = 1;
```

	p…		
0	3	898.3	5
1	4	982.15	6
2	5	832.63	7
3	11	1457.22	7
4	14	1450.5	6
5	17	1040.51	5
6	20	1792.63	7
7	28	1036.72	5
8	29	1031.11	7
9	30	1405.4	5
10	35	1039.58	6
11	36	879.37	4
12	37	1034.96	7
13	41	1074.63	3
14	43	615.07	5
15	46	1063.91	5

Rows: 552 Expand