УДК 51985

Кравчинський А.В.[1]
Наук.керівник: Бочаров Б.П.[2]

ЗАСТОСУВАННЯ ПЕРЕТВОРЕННЯ SQL

Мета цього експерімента - продемонструвати, як можна виконувати стандартні операції SQL, такі як з'єднання, об'єднання та його узагальнення, в машинному навчанні Microsoft Azure за допомогою модуля «Застосувати перетворення SQL». Використовуючи цей модуль, можна виконувати різні перетворення даних за допомогою SQL.

В цьому експерименті ми використовуємо три набори даних (клієнт ресторану, характеристика ресторану і рейтинги ресторану). Набори даних включають як числові, так і категоріальні характеристики.

На наступній діаграмі зображено загальний робочий процес експерименту (рис. 1):

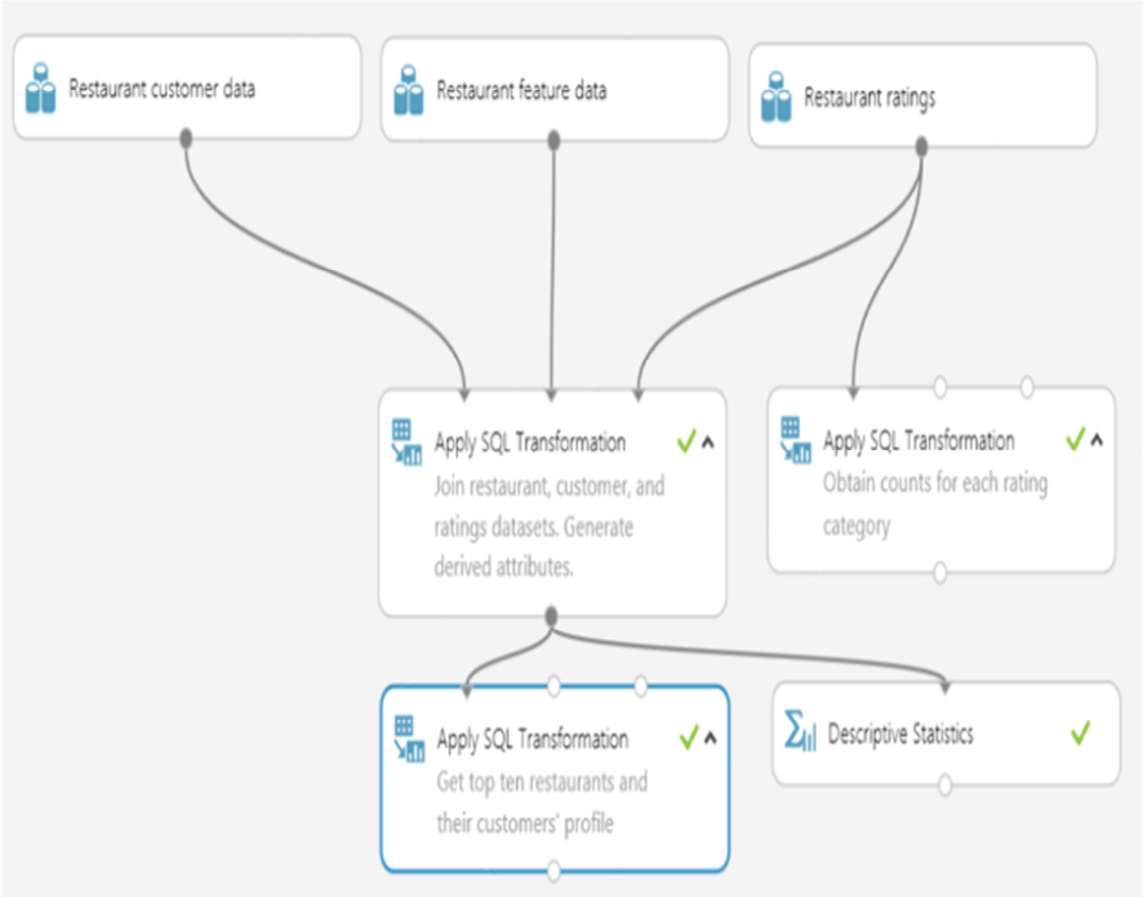

Рисунок 1 – процес експерименту

[1]Студент групи М КН 2020-1, ХНУМГ ім. О.М.Бекетова
[2] К.т.н., доцент кафедри КНтаІТ ХНУМГ ім. О.М.Бекетова

По-перше, було проведено просте дослідження набору даних рейтингів ресторанів і отримали кількість ресторанів для кожної рейтингової категорії. В цьому випадку в модуль «Застосувати перетворення SQL» було надано тільки один вхідний набір даних (рис. 2).

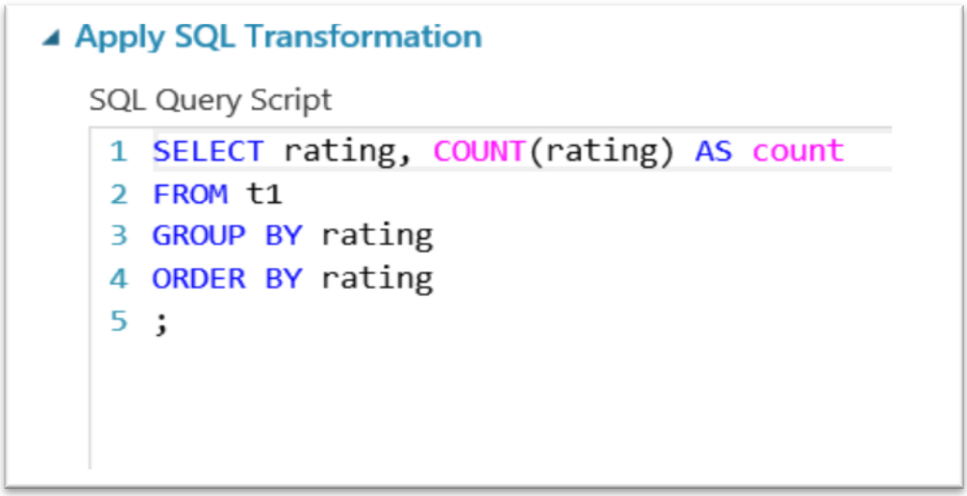

Рисунок 2 – Дослідження набору даних рейтингів ресторану

Результат зображено нижче (рис. 3):

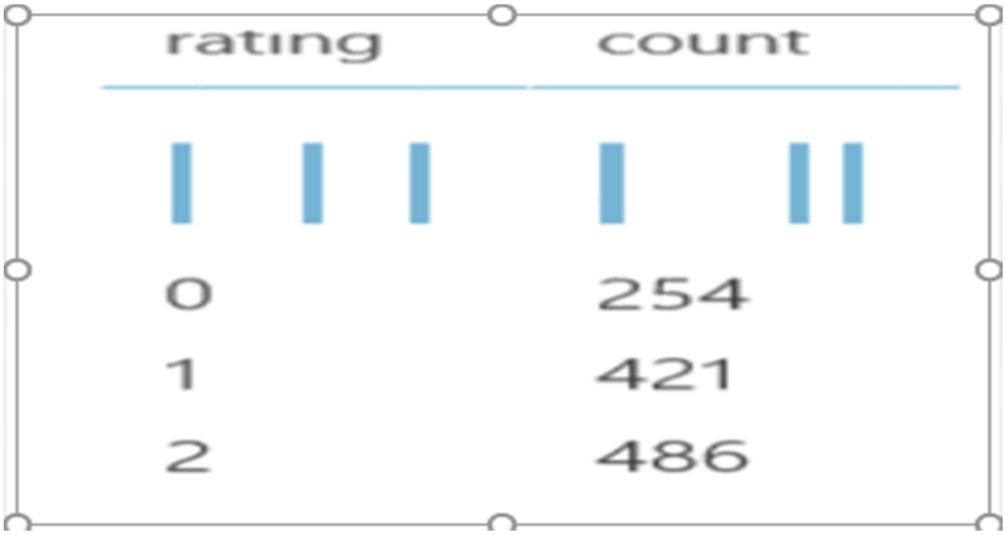

Рисунок 3 – Кількість ресторанів для кожної рейтингової категорії

По-друге, було зроблено внутрішнє з'єднання всіх трьох наборів даних в SQL на основі ідентифікатора користувача і ідентифікатора місця. Крім того, ми витягли такі атрибути, як Вік (у 2015 році), на основі Birth_year з даних про відвідувачів ресторану (рис. 4).

Apply SQL Transformation

SQL Query Script

```
SELECT
t3.userID AS userID,
t3.placeID AS placeID,
(2015-t1.birth_year) AS age,
weight,
height,
budget,
color,
activity,
religion,
personality,
interest,
hijos,
marital_status,
transport,
ambience,
dress_preference,
drink_level,
smoker,
abs(t1.latitude-t2.latitude) AS latitude_diff,
abs(t1.longitude-t2.longitude) AS longitude_diff,
city,
state,
country,
alcohol,
smoking_area,
dress_code,
accessibility,
price,
Rambience,
franchise,
area,
other_services,
rating
FROM t1, t2, t3
WHERE  t1.userID =t3.userID
AND t2.placeID=t3.placeID;
```

Рисунок 4 – Внутрішнє з'єднання трьох наборів даних

Результат зображено нижче (рис. 5):

userID	placeID	age	weight	height	budget	color	activity	religion	personality	interest	hijos	marital_status
U1001	132825	26	69	1.77	medium	black	student	none	thrifty-protector	variety	independent	single
U1001	132830	26	69	1.77	medium	black	student	none	thrifty-protector	variety	independent	single
U1001	135025	26	69	1.77	medium	black	student	none	thrifty-protector	variety	independent	single

Рисунок 5 – Результат SQL запроса

Нарешті, було використано об'єднання результатів трьох таблиць (зверху) і обчислено такі показники, як середній рейтинг і загальні оцінки, отримані для кожного ресторану (рис. 6).

Apply SQL Transformation

SQL Query Script

```
SELECT
placeID,
AVG(rating) AS Average_Rating,
COUNT(rating) AS Num_of_Ratings,
price,
Rambience,
smoking_area,
AVG(age) AS Customer_Average_Age,
AVG(weight) AS Customer_Average_Weight,
AVG(height) AS Customer_Average_Height
FROM t1
GROUP BY
placeID
ORDER BY Average_Rating DESC
LIMIT 10
;
```

Рисунок 6 – Обчислення середнього рейтингу і оцінки ресторанів

Десять кращих ресторанів, відсортованих за середньою оцінкою в порядку убування (на виході з модуля «Застосувати перетворення SQL»), перераховані нижче (рис. 7).

placeID	Average_Rating	Num_of_Ratings	price	Rambience	smoking_area	Customer_Average_Age	Customer_Average_Weight	Customer_Average_Height
132955	2	5	low	familiar	none	35.6	47.8	1.564
134986	2	8	high	familiar	none	33.625	67.125	1.65875
135034	2	5	medium	familiar	none	28	53.6	1.67
132922	1.833333	6	medium	familiar	permitted	25.333333	57.666667	1.628333
132755	1.8	5	medium	familiar	none	27.4	59	1.69
134976	1.75	4	low	familiar	none	30.25	60.25	1.6425
135013	1.75	4	low	familiar	none	24.25	70.75	1.6275
135074	1.75	4	high	familiar	section	28.75	54.5	1.665
135055	1.714286	7	high	familiar	section	25.571429	59.285714	1.594286
135075	1.692308	13	medium	familiar	none	29.538462	60.384615	1.686154

Рисунок 7 – Відображення десяти кращих ресторанів

ПЕРЕЛІК ПОСИЛАНЬ

1. Tan P-N., Steinbach M., Karpatne A. and Kumar V. “Introduction to data mining”. Pearson; 2nd edition. 2018. 864p.
2. Ben-Gan, Itzik. “T-SQL Fundamentals”. 3rd edition. 2016. 235p.