1. ***TẠO DATABASE SAMPLE ( NƠI LƯU TRỮ DỮ LIỆU GỐC )***

* Chuột phải -> Import Flat File -> Chọn file.csv

A computer screen with a white box

AI-generated content may be incorrect.

* Dữ liệu gồm 13 trường & 705 dòng.

A screenshot of a computer

AI-generated content may be incorrect.

1. ***TẠO STAGING DATABASE ( NƠI XỬ LÝ DỮ LIỆU )***

- Chọn database SAMPLE cần nhân bản. Chọn Task >> Copy Database.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

1. ***LÀM SẠCH DỮ LIỆU LẦN 1***

Làm sạch nhẹ để đảm bảo dữ liệu có thể dùng phân tích, chưa biến đổi sâu.

* Tạo table GIAODICH\_CLEAN\_1 (cùng cấu trúc với table GIAODICH).

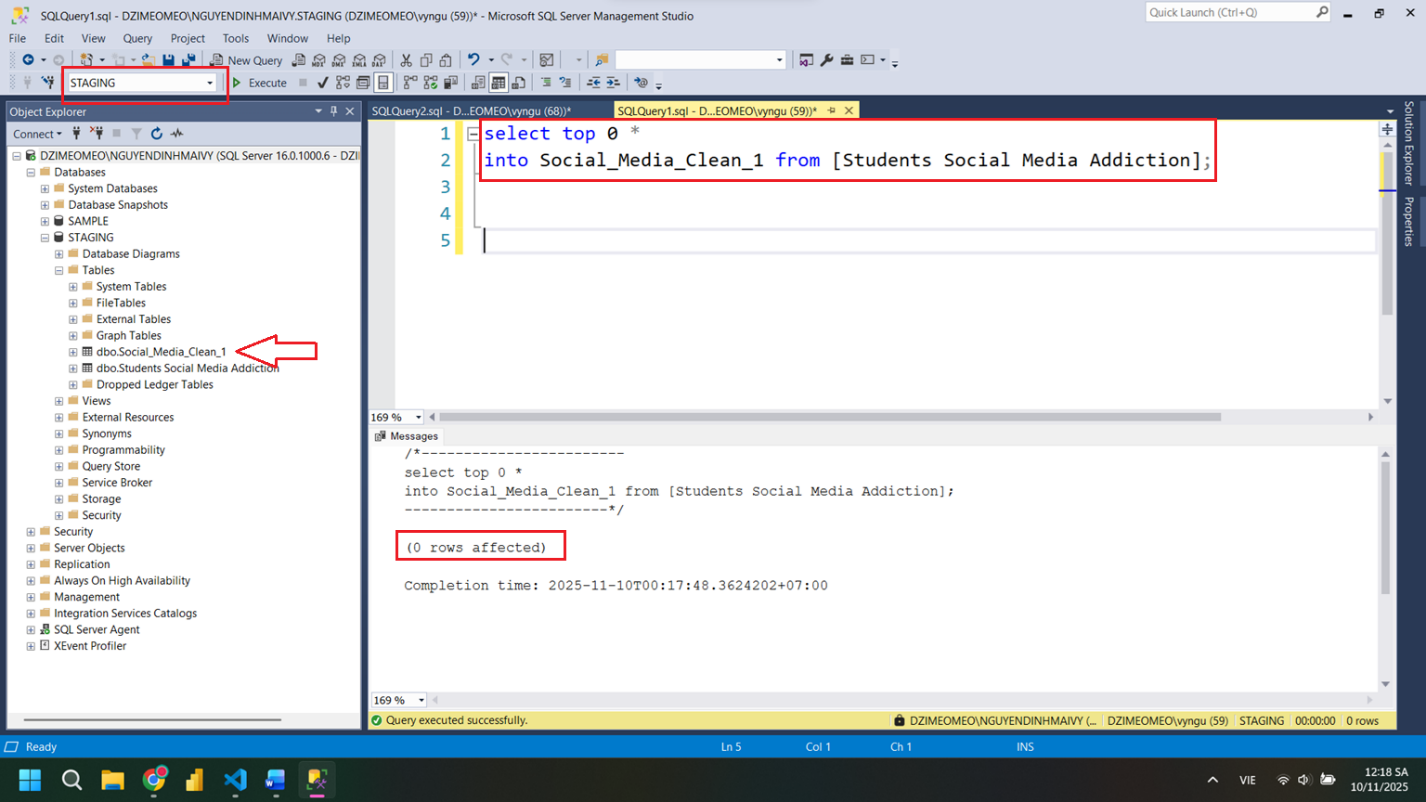
***Dùng công thức:***

* Chọn database Staging

select top 0 \* *[top 0 giúp tạo bảng mới với cấu trúc cột giống hệt mà không copy dữ liệu.]*

into Social\_Media\_Clean\_1

from [Students Social Media Addiction];



* Tiếp theo thêm vào bảng Clean\_1 dữ liệu gốc và xử lý lại theo các bước sau:

1. ***Loại bản ghi trùng:*** theo Student\_ID
2. ***Loại bản ghi có giá trị null / trống*** ở các cột chính (Student\_ID, Age, Gender, Country)
3. ***Chuẩn hóa text:*** xóa khoảng trắng, viết hoa chữ cái đầu nếu cần.
4. ***Chuyển đổi kiểu dữ liệu:***

Age, Mental\_Health\_Score, Addicted\_Score, Conflicts\_Over\_Social\_Media → INT

Avg\_Daily\_Usage\_Hours, Sleep\_Hours\_Per\_Night → FLOAT

**Dùng công thức:**

INSERT INTO Social\_Media\_Clean\_1

SELECT DISTINCT

TRY\_CAST(Student\_ID AS INT) AS Student\_ID,

TRY\_CAST(Age AS INT) AS Age,

TRIM(CAST(Gender AS NVARCHAR(50))) AS Gender,

TRIM(CAST(Academic\_Level AS NVARCHAR(100))) AS Academic\_Level,

TRIM(CAST(Country AS NVARCHAR(100))) AS Country,

TRY\_CAST(Avg\_Daily\_Usage\_Hours AS FLOAT) AS Avg\_Daily\_Usage\_Hours,

TRIM(CAST(Most\_Used\_Platform AS NVARCHAR(100))) AS Most\_Used\_Platform,

TRIM(CAST(Affects\_Academic\_Performance AS NVARCHAR(10))) AS Affects\_Academic\_Performance,

TRY\_CAST(Sleep\_Hours\_Per\_Night AS FLOAT) AS Sleep\_Hours\_Per\_Night,

TRY\_CAST(Mental\_Health\_Score AS INT) AS Mental\_Health\_Score,

TRIM(CAST(Relationship\_Status AS NVARCHAR(50))) AS Relationship\_Status,

TRY\_CAST(Conflicts\_Over\_Social\_Media AS INT) AS Conflicts\_Over\_Social\_Media,

TRY\_CAST(Addicted\_Score AS INT) AS Addicted\_Score

FROM [Students Social Media Addiction]

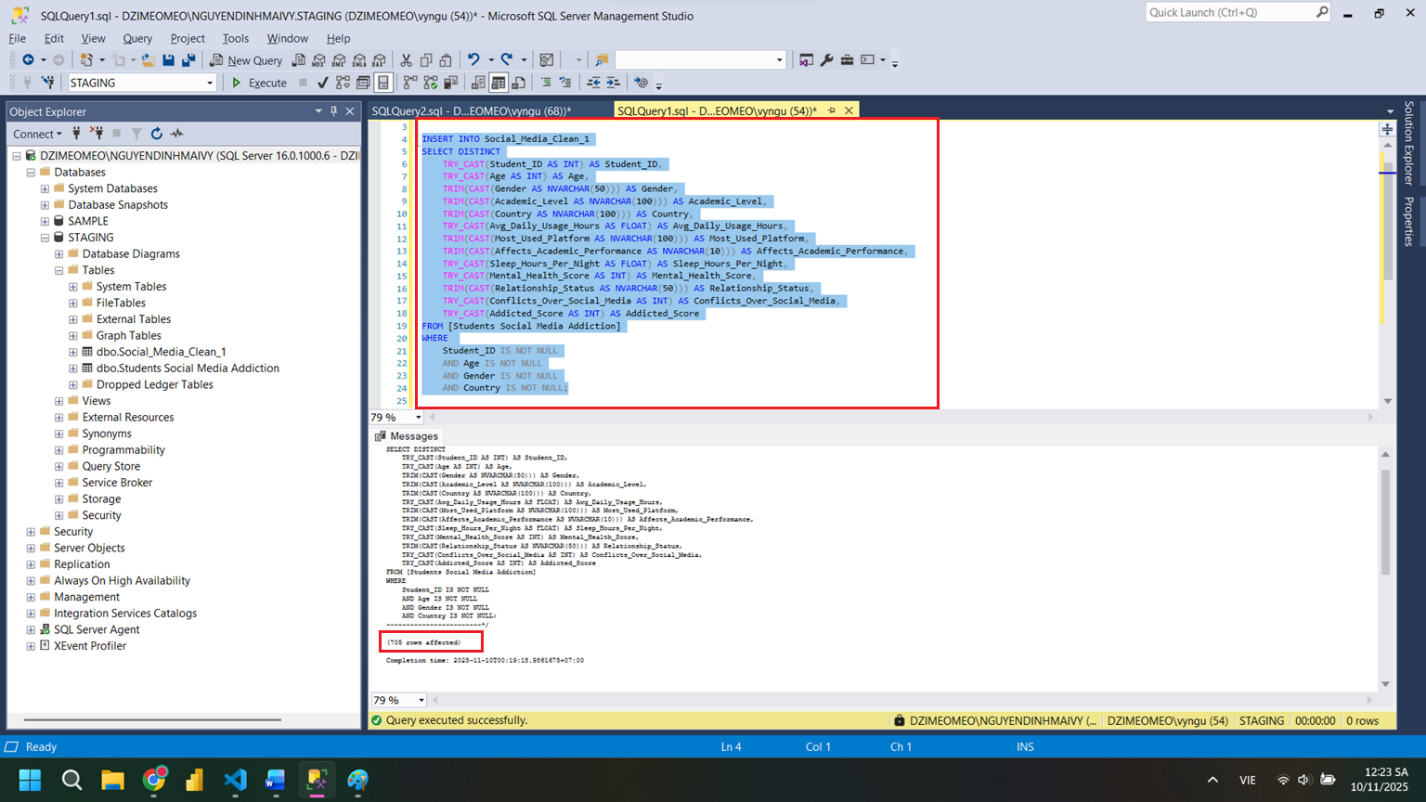
WHERE

Student\_ID IS NOT NULL

AND Age IS NOT NULL

AND Gender IS NOT NULL

AND Country IS NOT NULL;



A computer screen shot of a computer screen

AI-generated content may be incorrect.

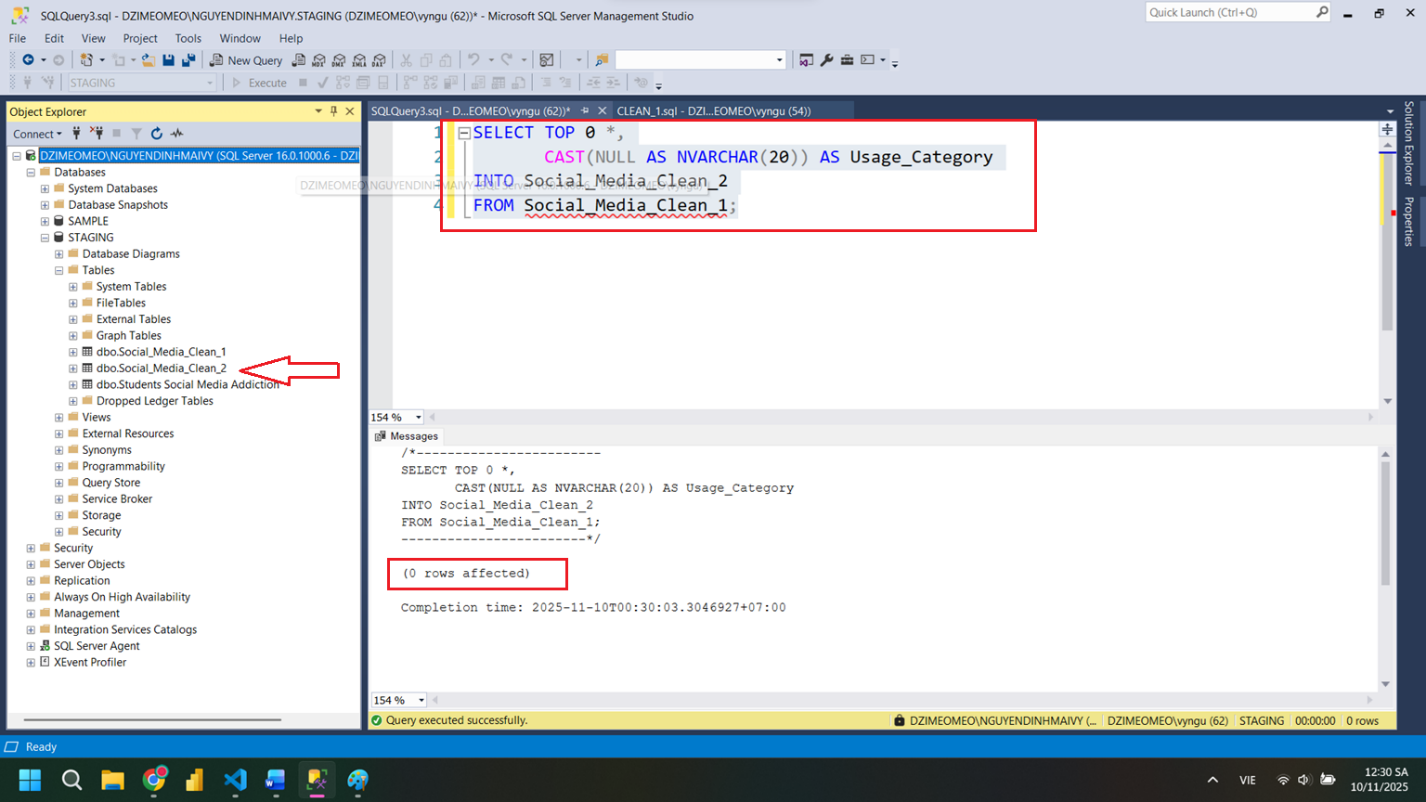
Kết quả: 13 cột & 705 dòng

1. ***LÀM SẠCH DỮ LIỆU FINAL***

Chuẩn hóa các giá trị:

| **Cột** | **Quy tắc làm sạch / chuyển đổi** |
| --- | --- |
| Gender | Chuẩn hóa Male / Female (viết hoa chữ cái đầu) |
| Academic\_Level | Chuyển về dạng viết hoa chữ đầu (High School, Undergraduate, Graduate) |
| Country | Xóa khoảng trắng, viết hoa chữ đầu |
| Relationship\_Status | Chuẩn hóa (Single, In Relationship, Complicated) |
| Avg\_Daily\_Usage\_Hours | Phân loại: |

* Tạo bảng Clean\_Final có cấu trúc giống Clean\_1 ( dữ liệu trống )



* Làm sạch và thêm dữ liệu từ Clean\_1

***Dùng công thức:***

INSERT INTO Social\_Media\_Clean\_Final

SELECT

Student\_ID,

Age,

UPPER(LEFT(Gender,1)) + LOWER(SUBSTRING(Gender,2,LEN(Gender))) AS Gender,

UPPER(LEFT(Academic\_Level,1)) + LOWER(SUBSTRING(Academic\_Level,2,LEN(Academic\_Level))) AS Academic\_Level,

UPPER(LEFT(Country,1)) + LOWER(SUBSTRING(Country,2,LEN(Country))) AS Country,

Avg\_Daily\_Usage\_Hours,

Most\_Used\_Platform,

CAST(Affects\_Academic\_Performance AS INT) AS Affects\_Academic\_Performance,

Sleep\_Hours\_Per\_Night,

Mental\_Health\_Score,

CASE

WHEN Relationship\_Status LIKE '%single%' THEN 'Single'

WHEN Relationship\_Status LIKE '%relationship%' THEN 'In Relationship'

WHEN Relationship\_Status LIKE '%complicated%' THEN 'Complicated'

ELSE 'Unknown'

END AS Relationship\_Status,

Conflicts\_Over\_Social\_Media,

Addicted\_Score,

CASE

WHEN Avg\_Daily\_Usage\_Hours > 6 THEN 'High Usage'

WHEN Avg\_Daily\_Usage\_Hours BETWEEN 3 AND 6 THEN 'Medium Usage'

ELSE 'Low Usage'

END AS Usage\_Category

FROM Social\_Media\_Clean\_1;

A computer screen shot of a computer screen

AI-generated content may be incorrect.

A computer screen shot of a computer screen

AI-generated content may be incorrect.

Kết quả: 705 dòng & 14 cột

1. **TẠO DATABASE SPEND ( DỮ LIỆU ĐÃ SẴN SÀNG ĐỂ PHÂN TÍCH )**

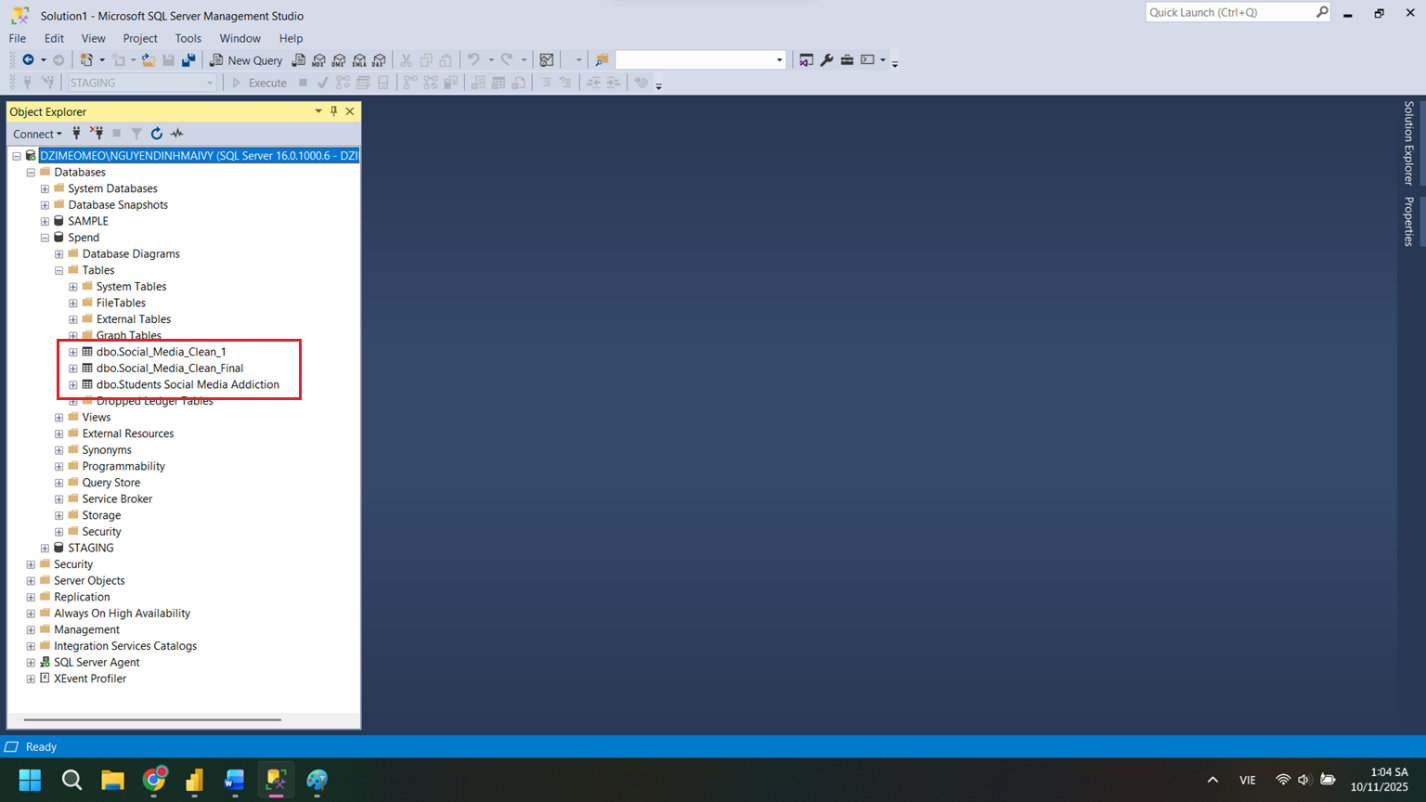
* Copy database Staging -> đặt tên Spend.

A computer screen with a message box

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.



* Kết nối vào powerbi

A screenshot of a computer

AI-generated content may be incorrect.

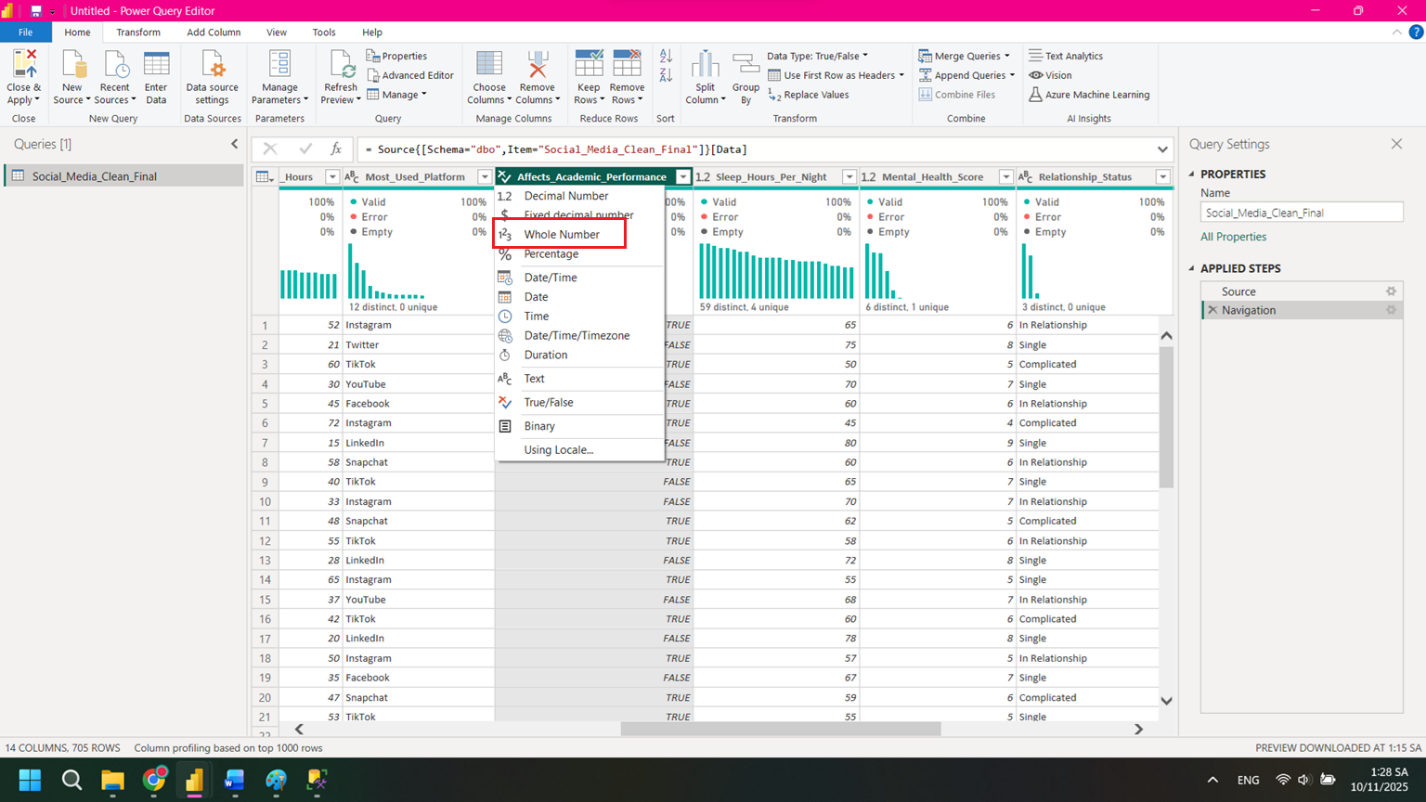
1. ***CHUẨN HÓA DỮ LIỆU***

* Chọn bảng Final -> Transform

A screenshot of a computer

AI-generated content may be incorrect.

* Định dạng lại dữ liệu cột [Affects\_Academic\_Performance]



* Chuột phải cột Platfrom -> Group by

A screenshot of a computer

AI-generated content may be incorrect.