Đồ án cuối khóa - Phân tích dữ liệu

Mục tiêu: Phân tích mức độ quan tâm của người dùng về tất cả các app có trên Google Play Store

Yêu cầu: Thể hiện trức quan các yếu tố về lượt tải, lượt đánh giá tỷ lệ đánh giá, sao đánh giá

Hướng giải quyết: Xây dựng 3 dashboard mỗi dashboard thể hiện mối tương quan của các yêu tố trên với một đặt điểm thông kê riêng như là loại ứng dụng, nội dung đánh giá, sao đánh giá

Cơ sở kiến thức: Cấu trúc cơ sở dữ liệu, ETL, trực quan hoá dữ liệu và nhiều kiến thức có liên quan khác

Nguồn dữ liệu: Tệp dữ liều được dùng đễ phân tích và báo cáo được thu thập từ trang keggle và có đường link: [Google Play Store Apps | Kaggle](https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps)

Tỗng quan về tệp dữ liệu:

Tệp dữ liệu có kích thước 676,46MB

Số lượng cột là 24 cột

Số lượng dòng là 2312944 dòng

Có các cột lần lượt là:

App Name: Tên ứng dụng

App Id: ID ứng dụng

Category: Loại ứng dụng

Rating: Số sao đánh giá

Rating Count: Số lượt đánh giá

Installs: Số lượt tải làm tròn

Minimum Installs: Số lượt tải

Maximum Installs: Số lượt tải chính xát

Free: Ứng dụng có trả phí hay không

Price: Giá của ứng dụng

Currency: Đơn vị tiền tệ thanh toán

Size: Kích thướt ứng dụng

Minimum Android: Yêu cầu phần mềm

Developer Id: ID nhà phát triển

Developer Website: Website nhà phát triển

Developer Email: Email nhà phát triển

Released: Ngày phát hành

Privacy Policy: Chính sách bảo mật từ nhà phát triển

Last Updated: Ngày cập nhật gần nhất

Content Rating: Giới hạn độ tuổi người dùng

Ad Supported: Hỗ trợ quãng cáo bên trong ứng dụng

In app purchases: Mua hàng bên trong ứng dụng

Editor Choice: Được đánh giá và lựa chọn của biên tập

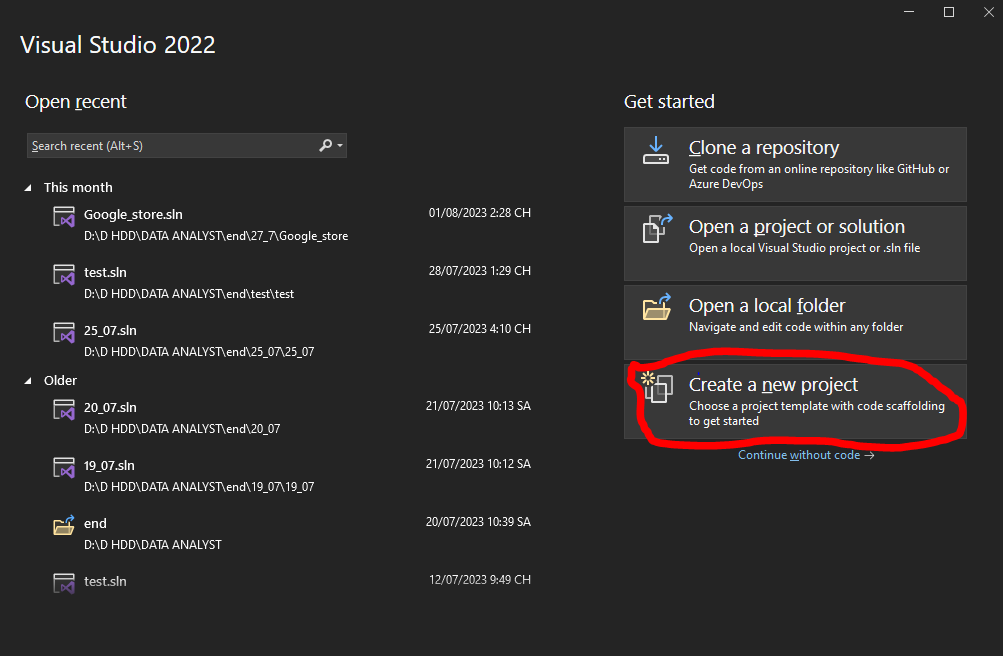
Sử dụng SISS để tiến hành ETL tệp dữ liệu trên:

Các bước tiến hành cụ thể như sau:

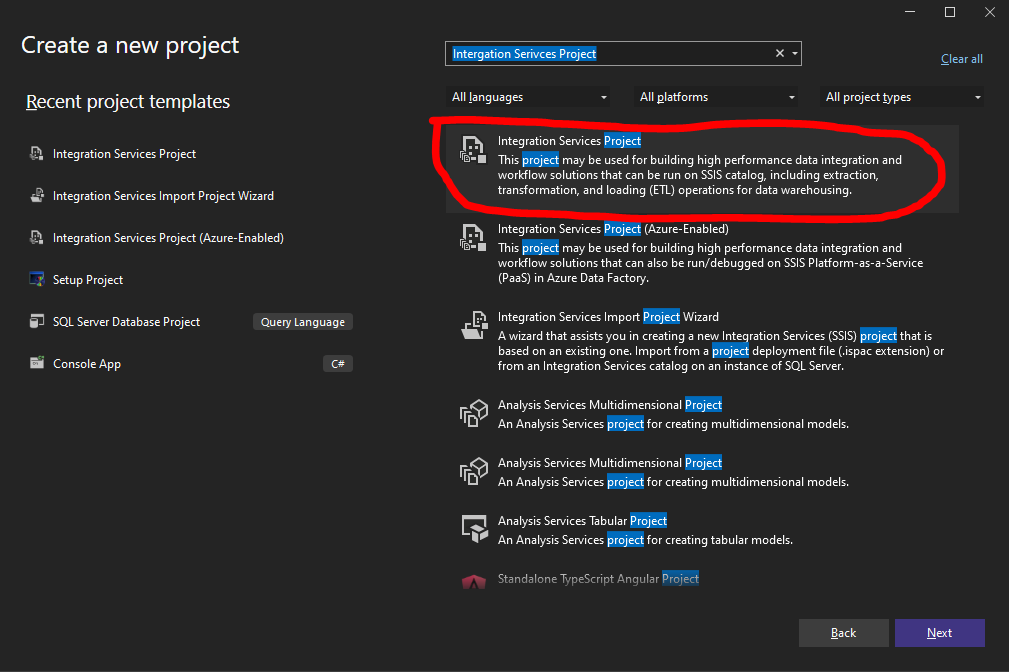
Bước 1:

Mỡ Microsoft Visual Studio 2022

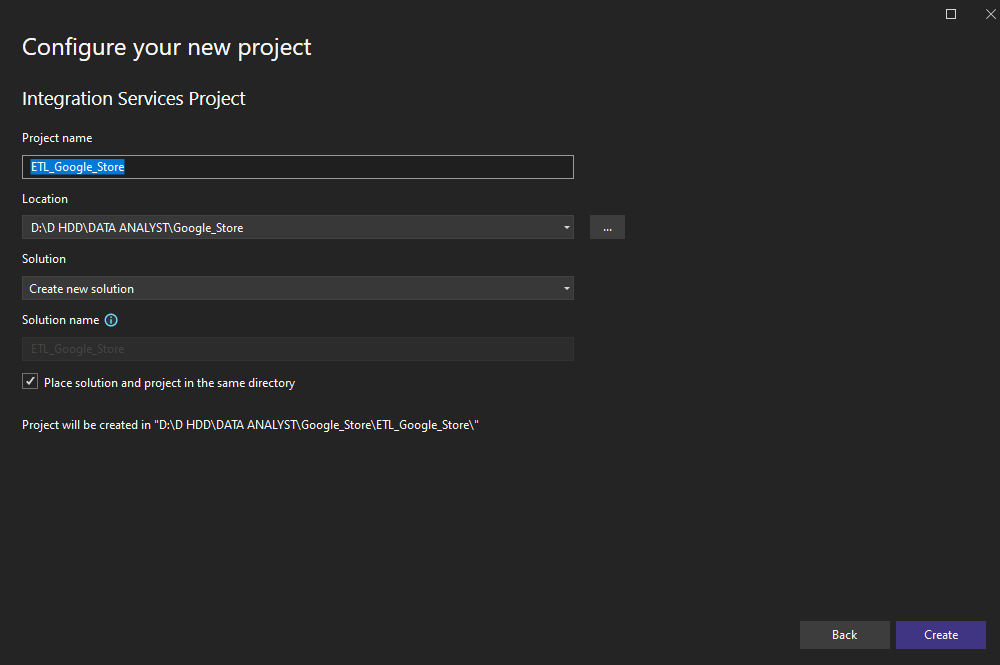
Tạo New Project nhấn vào Create a new project



Tìm Intergation Serivces Project trong phần tìm kiếm và click vào Intergation Serivces Project



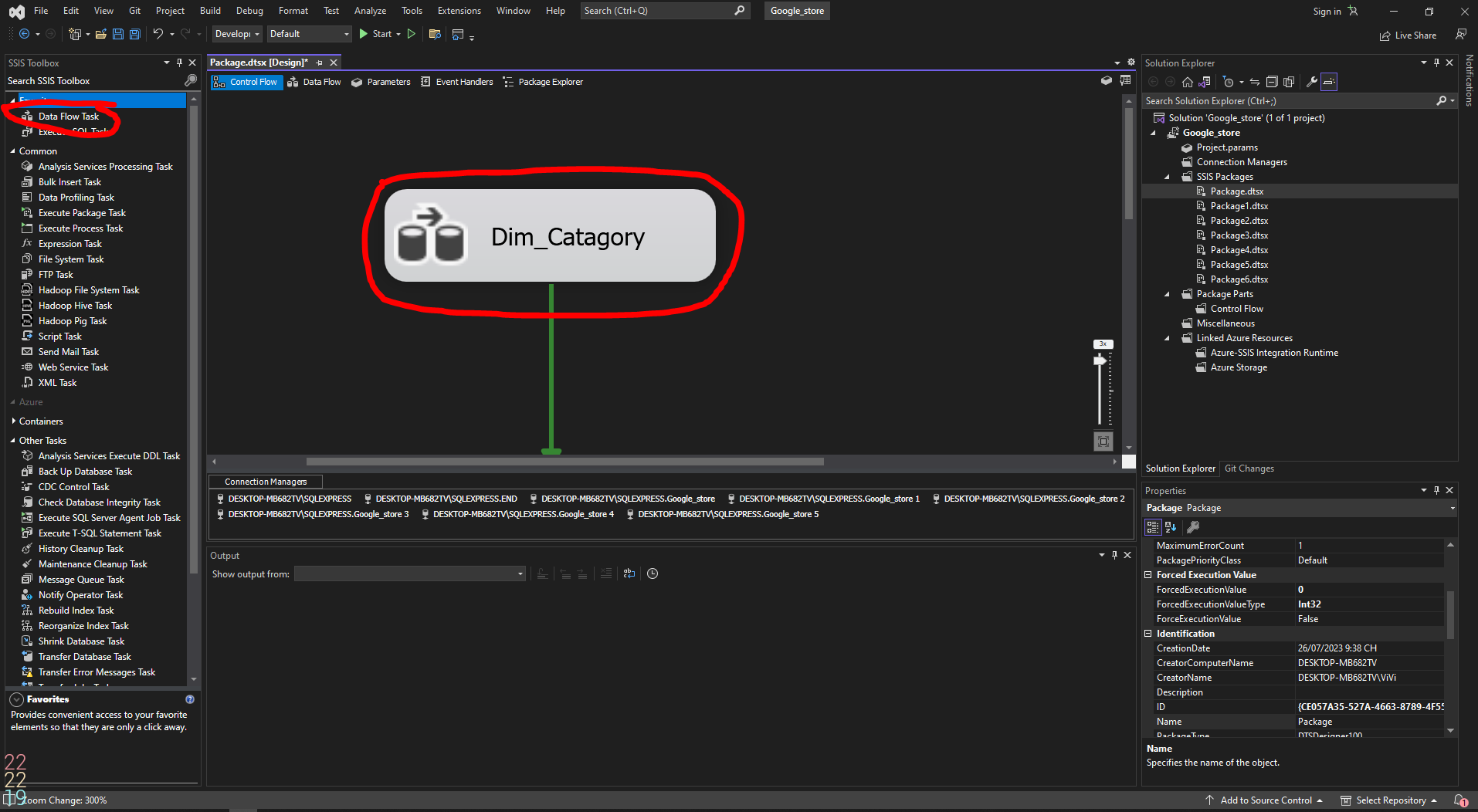
Đặt tên project, chọn nơi lưu project và nhấn create



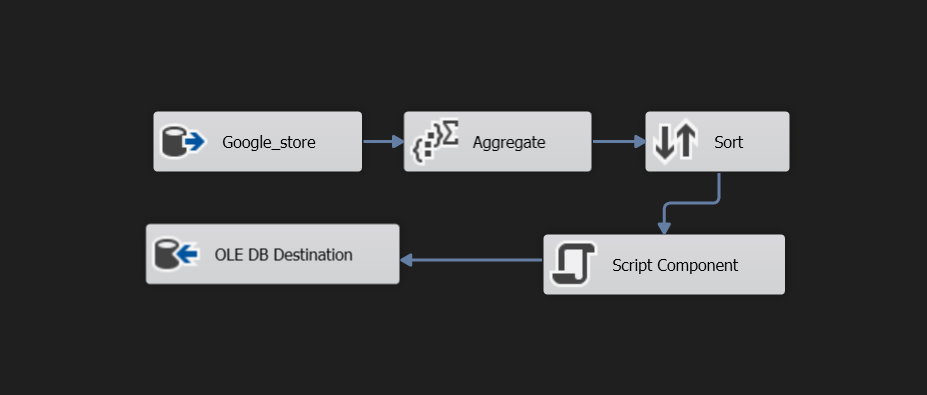
Bước 2:

Tạo các Data Flow Task

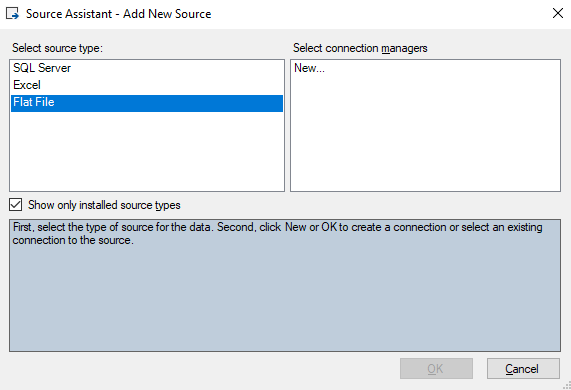
Tạo Data Flow Task thực hiện task tạo Dim\_Catagory



Nhấn dúp vào Toolbox Data Flow Task và tạo toolbox



Tạo toolbox Source Assistant đễ truyền file dữ liệu vào

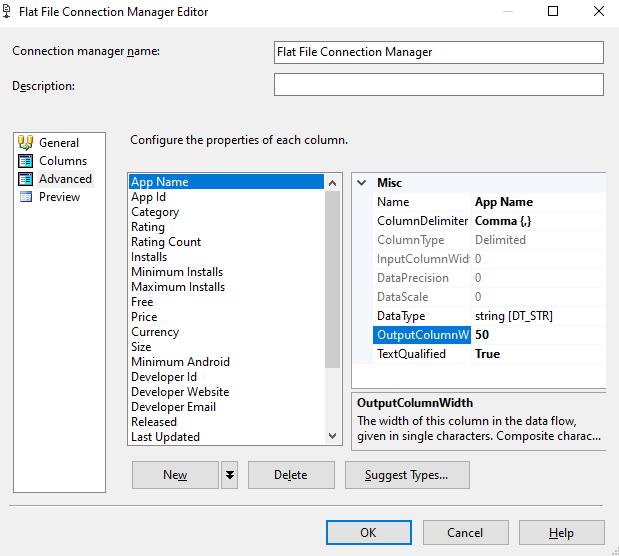


Chọn Flat File rồi chọn New xong nhấn OK. Khi hoàn thành cửa sỗ Flat File Connection Manager Editor hiện lên.

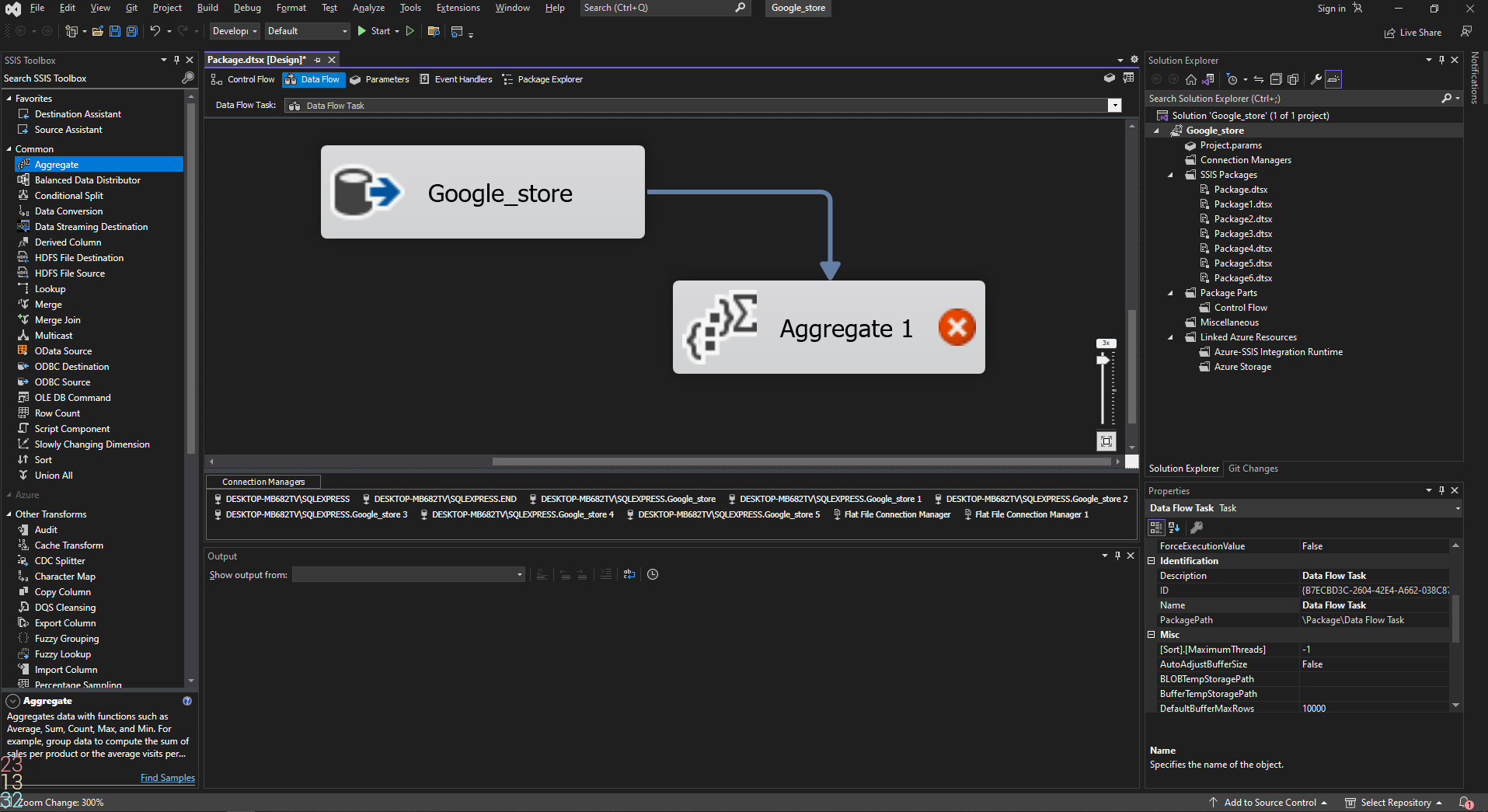
Đặt lại tên cho connection là Google\_store và chọn file dữ liễu đã chuẩn bị sẵn từ trước bằng nhấn vào Browse, chú ý phần tích chọn column names in first data row.



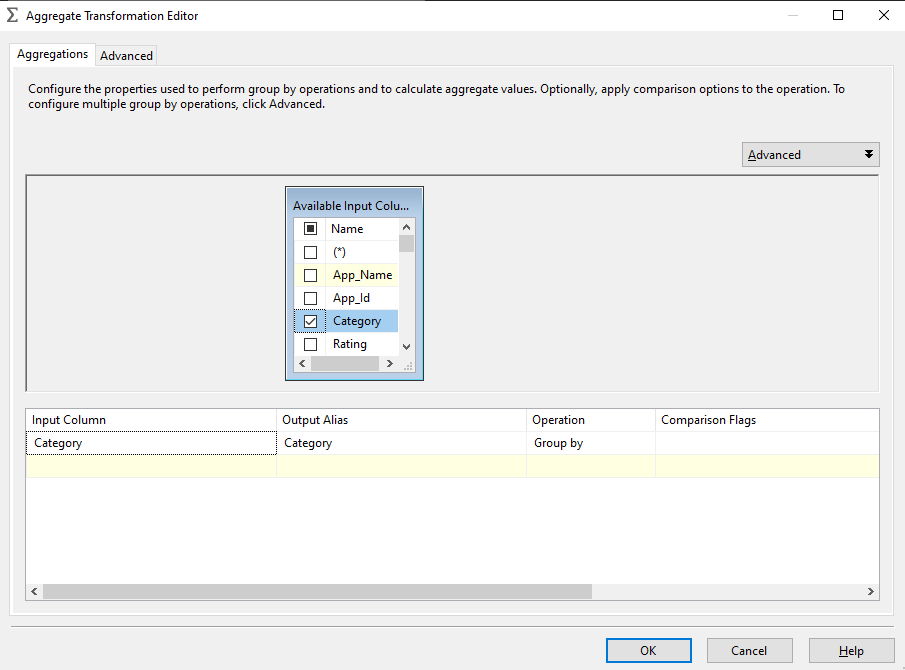
Click vào tab advanced và chọn DataType phù hợp cho từng cột, sao đó chọn tab Preview đễ kiễm tra lại dữ liệu, nếu dữ liệu không có phát sinh vấn đề thì nhấn OK.



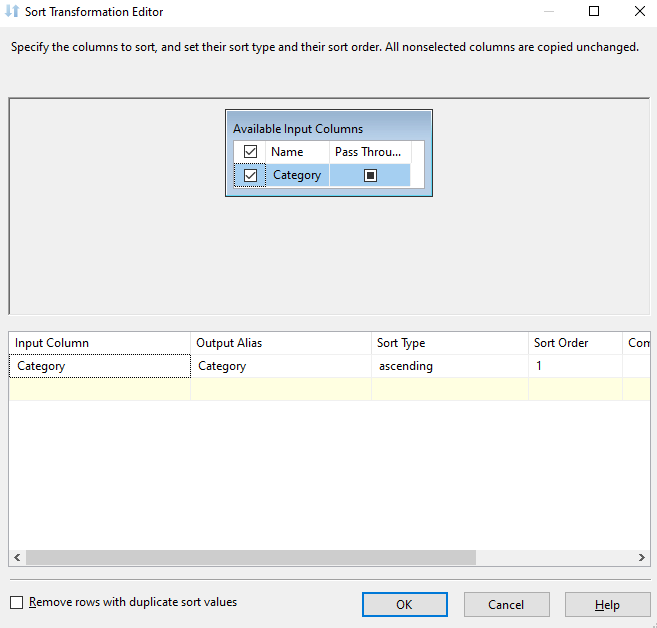
Tiếp theo tiến hành Group By cột Category bằng cách toạ toolbox Aggregate, sao đó nối khối Google\_store với khối Aggregate vừa tạo.



Nhấn đúp vào khối Aggragate đễ mớ hột thoại Aggregate Transformation Editor và chọn cột Category, sau đó chọn OK

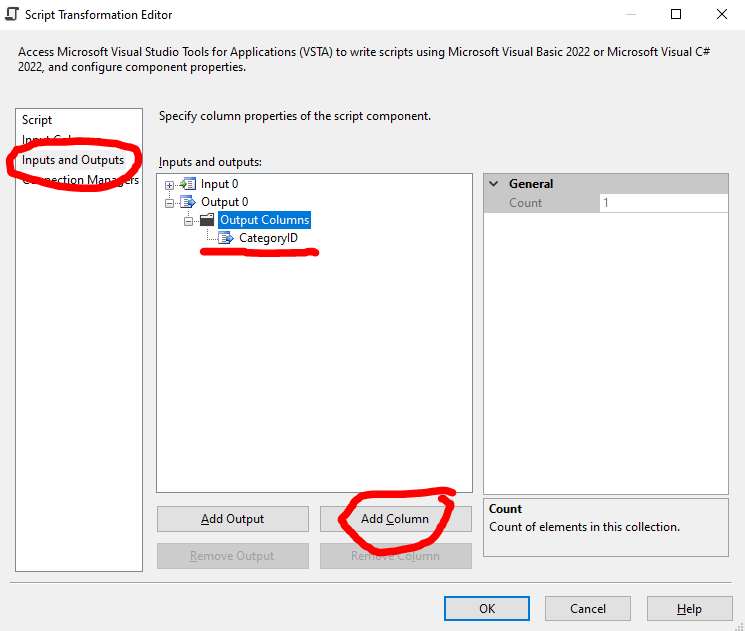


Chọn toolbox Sort và kết nối với toolbox Aggregate sau đó tích chọn cột Category

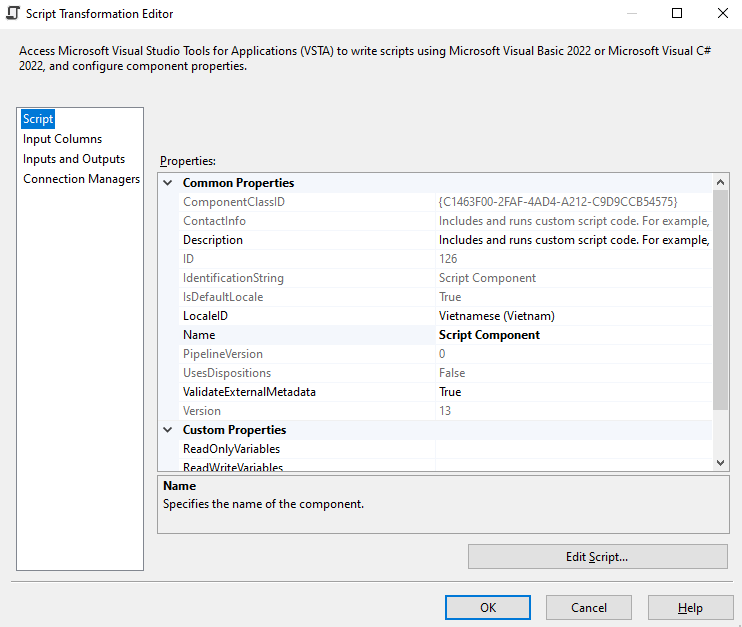


Tạo cột CategoryID bằng cách thêm toolbox Script Component và kết nối với toolbox Sort sau đó click đúp vào toolbox Script Component đễ mỡ hộp thoại Script Transformation Editor.

Tại tab Inputs and Outputs và chọn mục Output Columns sau đó nhấn Add Column và đặt tên cột là CategoryID.



Tại tab Script click vào Edit Script

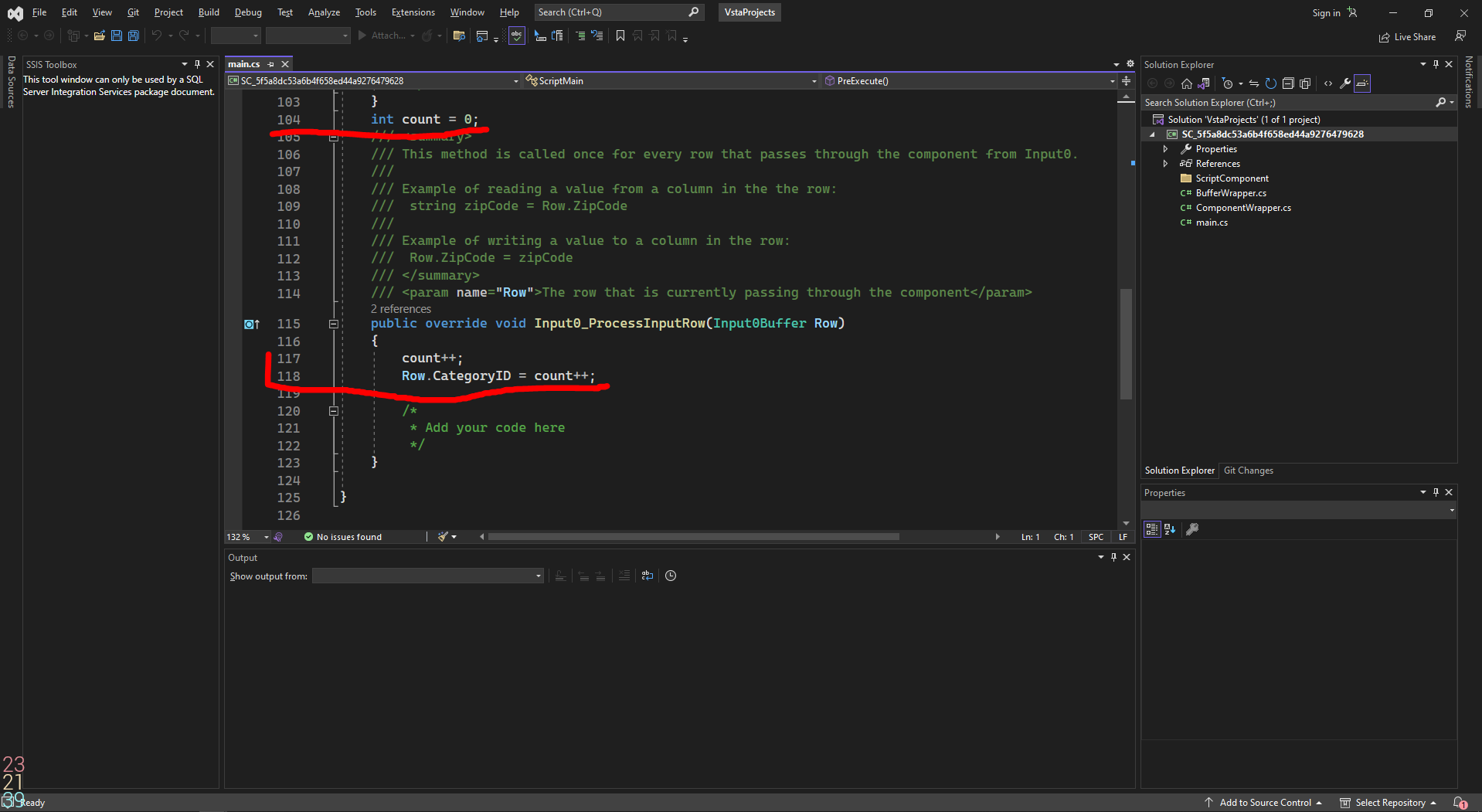


Tại cửa sổ VstaProjects, sửa lại chương trình bằng cách thêm vào các dòng các dòng code:

int count = 0;

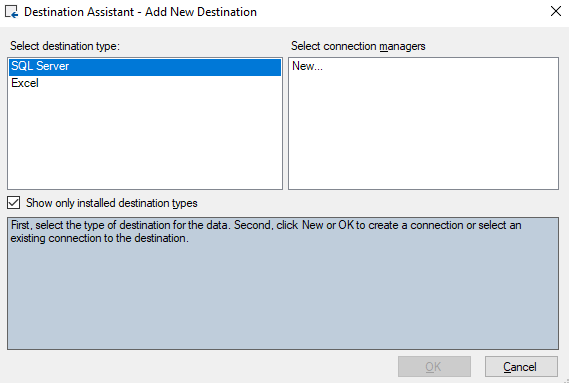
count ++;

Row.CategoryID = count++;

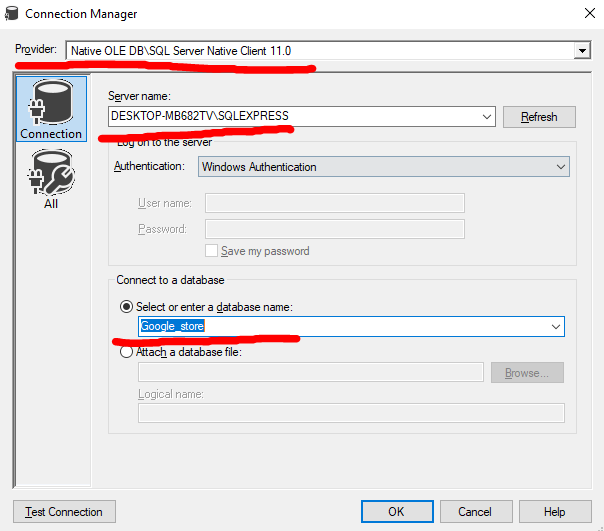


Lưu lại chương trinhg và đóng cửa sỗ, sau đó nhấn OK tại hộp thoại Script Transformation Editor.

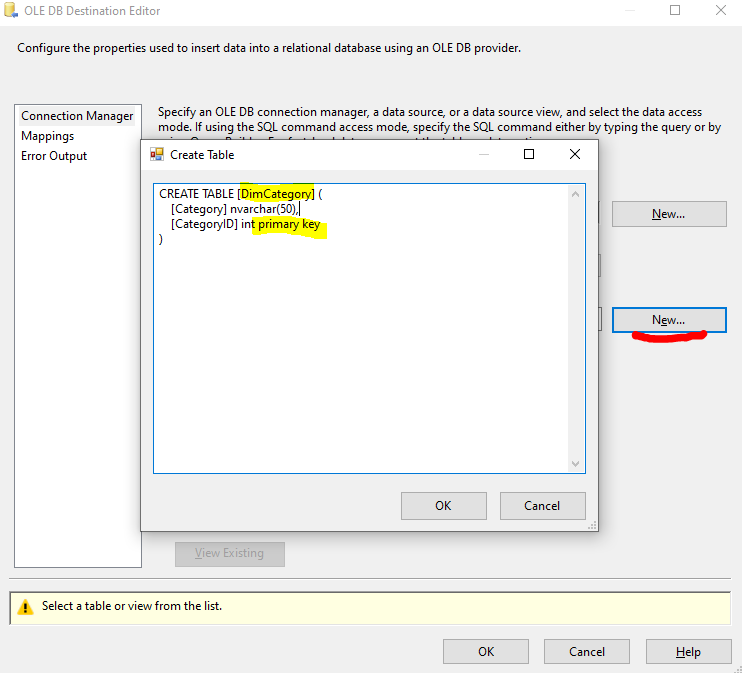
Nhấn đúp vào Destination Assistant sau đó hộp thoại Destination Assistant, chọn SQL Server và chọn New sau đó nhấn OK



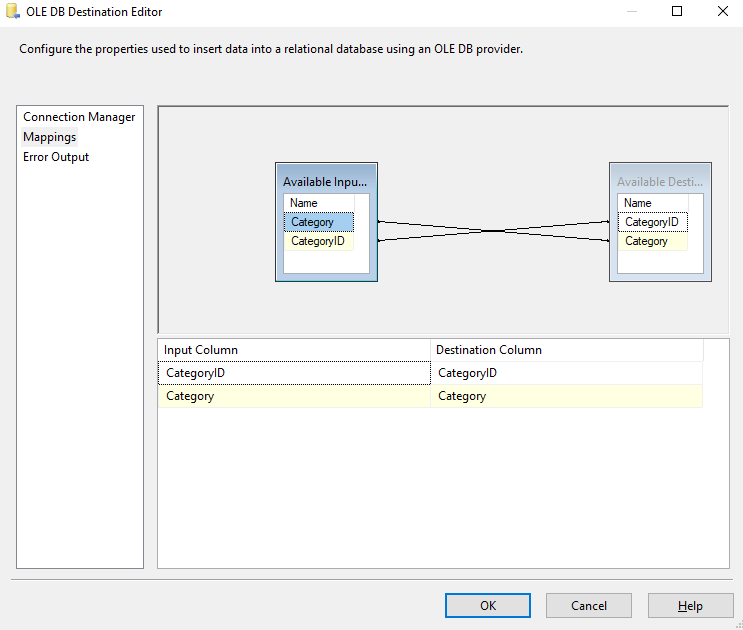
Tại hộp thoại Connection Manager lựa chọn Provider là SQL Server Native Client 11.0, chọn Server và database phù hợp ở những phần được đánh dấu, sau đó nhấn OK



Kết nối Script Component với OLE DB Destination sau đó nhấn đúp vào OLE DB Destination đễ mỡ hộp thoại OLE DB Destination Editor chọn New… đễ tạo bảng mới trong database và đặt lại tên bảng và primary key sau đó nhấn OK.



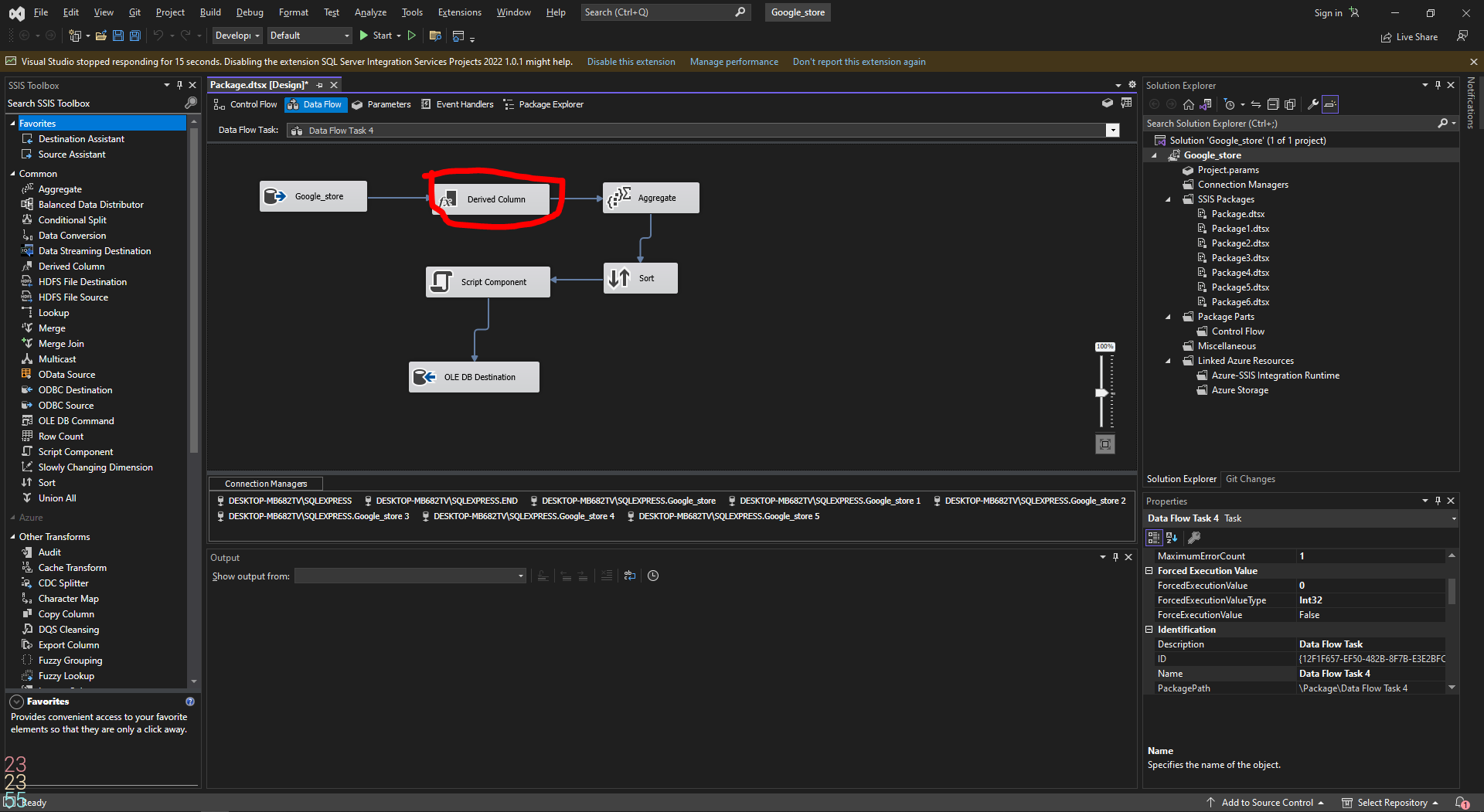
Tại tab Mappings kiểm tra lại các mapping có chính xát chưa nhấn OK đễ đóng hộp thoại



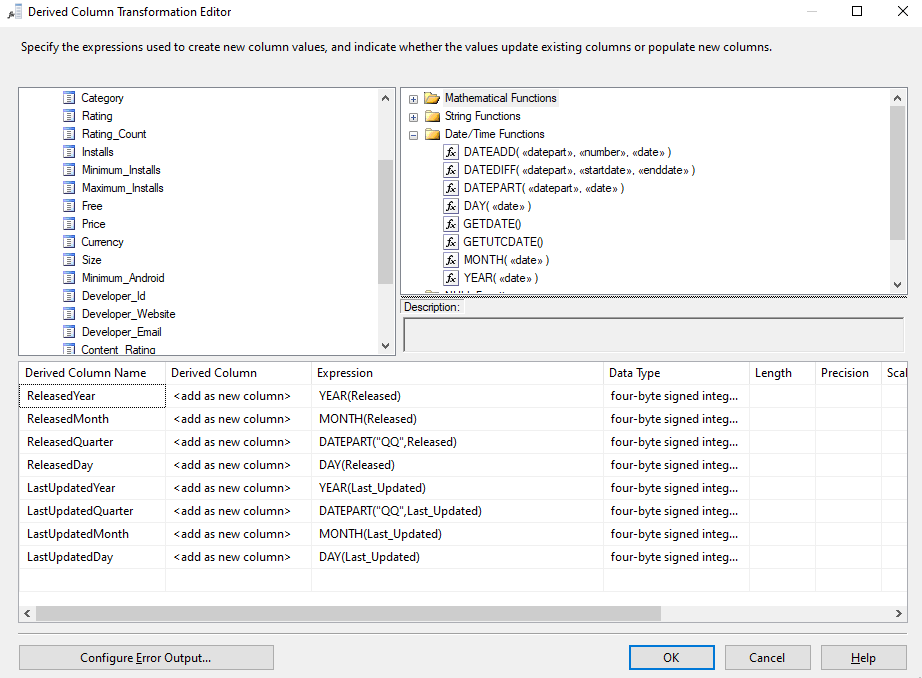
Bước 3

Thực hiện thao tác tương tự đễ tạo DimRating, DimPrice, DimSize, DimDate

Lưu ý khi tạo DimDate:

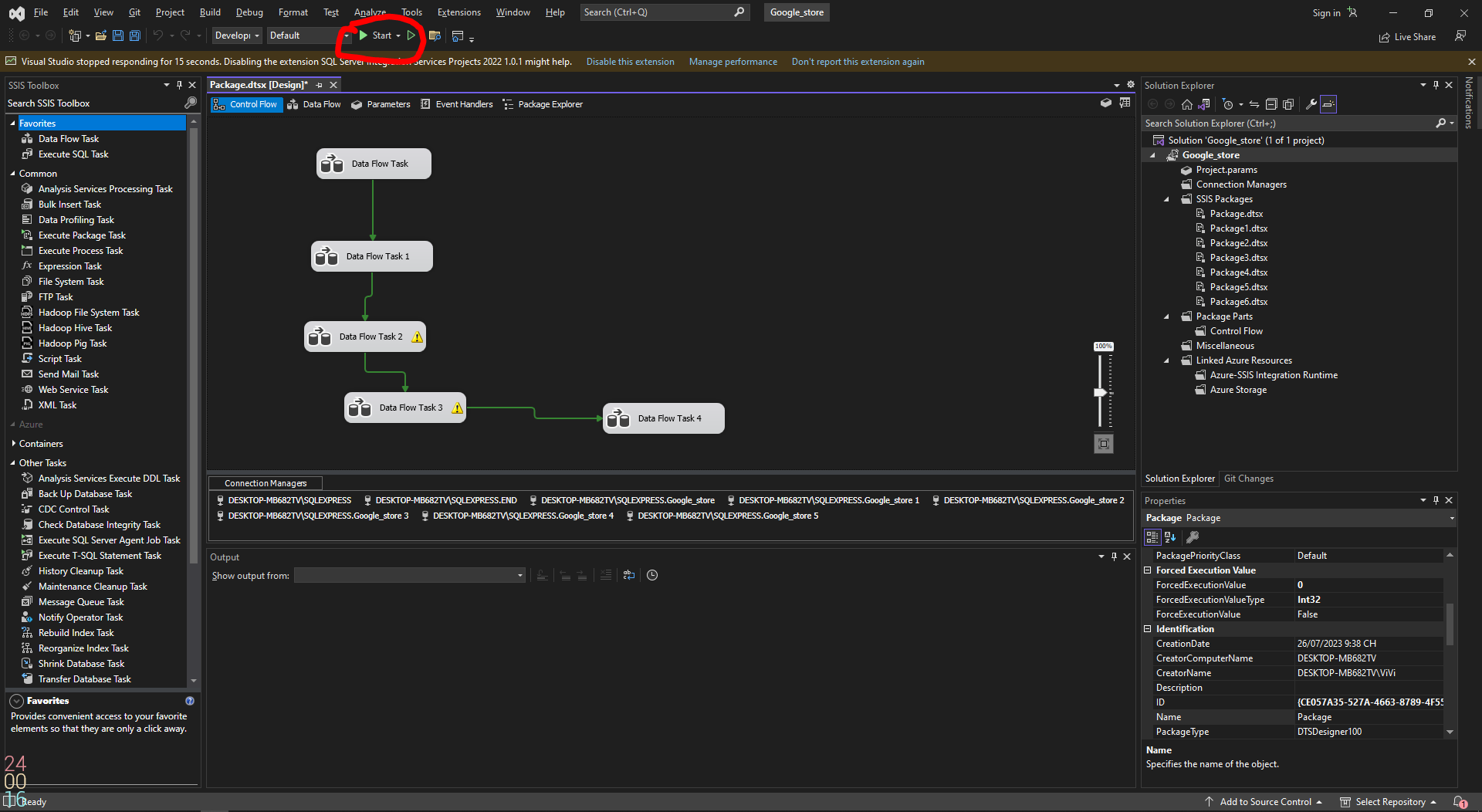


Tạo thêm khối Derived Column đễ tách giá trị ngày tháng năm quý ra thành từng cột bằng cách sử dụng các lệnh như hình



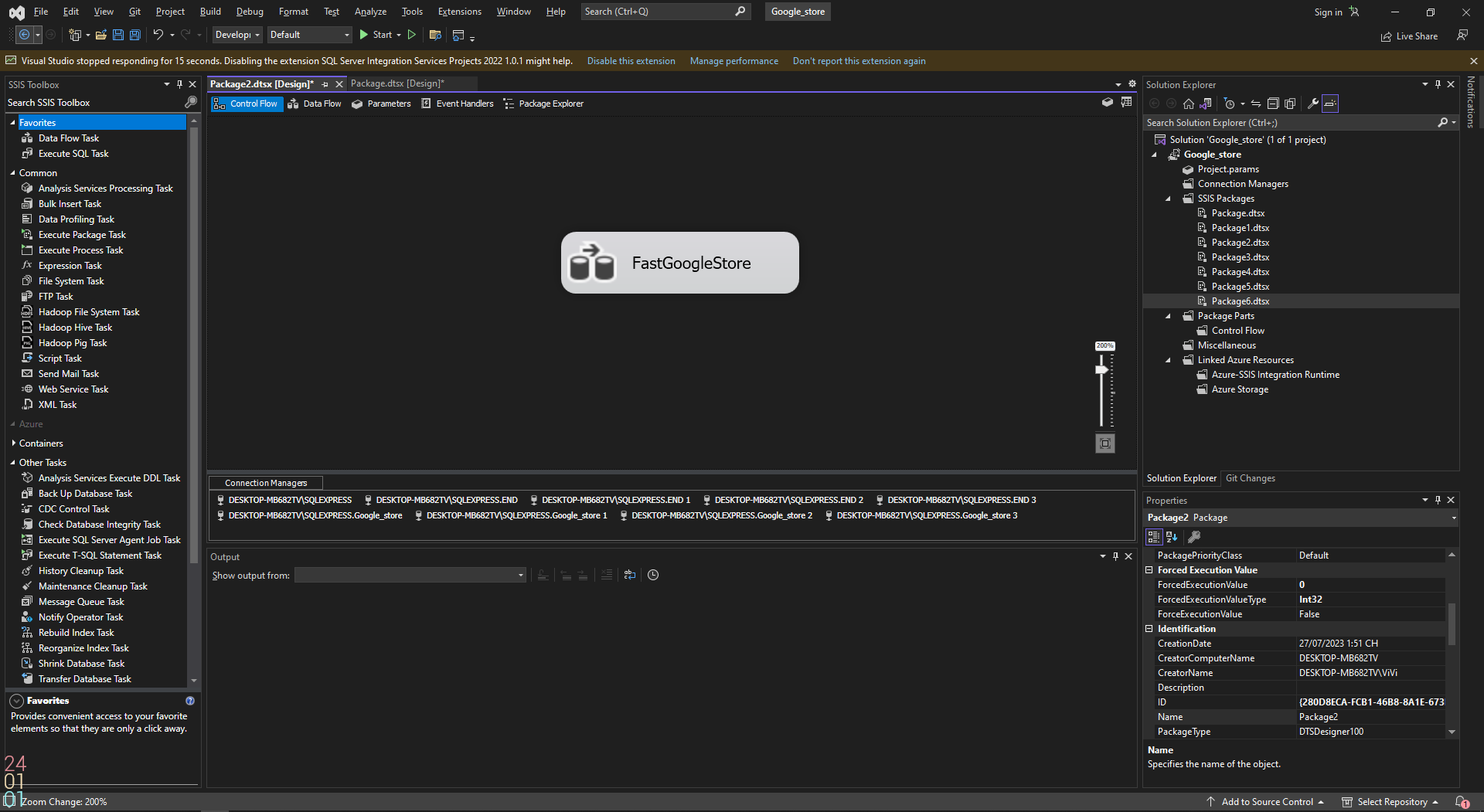
Bước 4:

Chạy Project ETL bằng cách nhấn Start

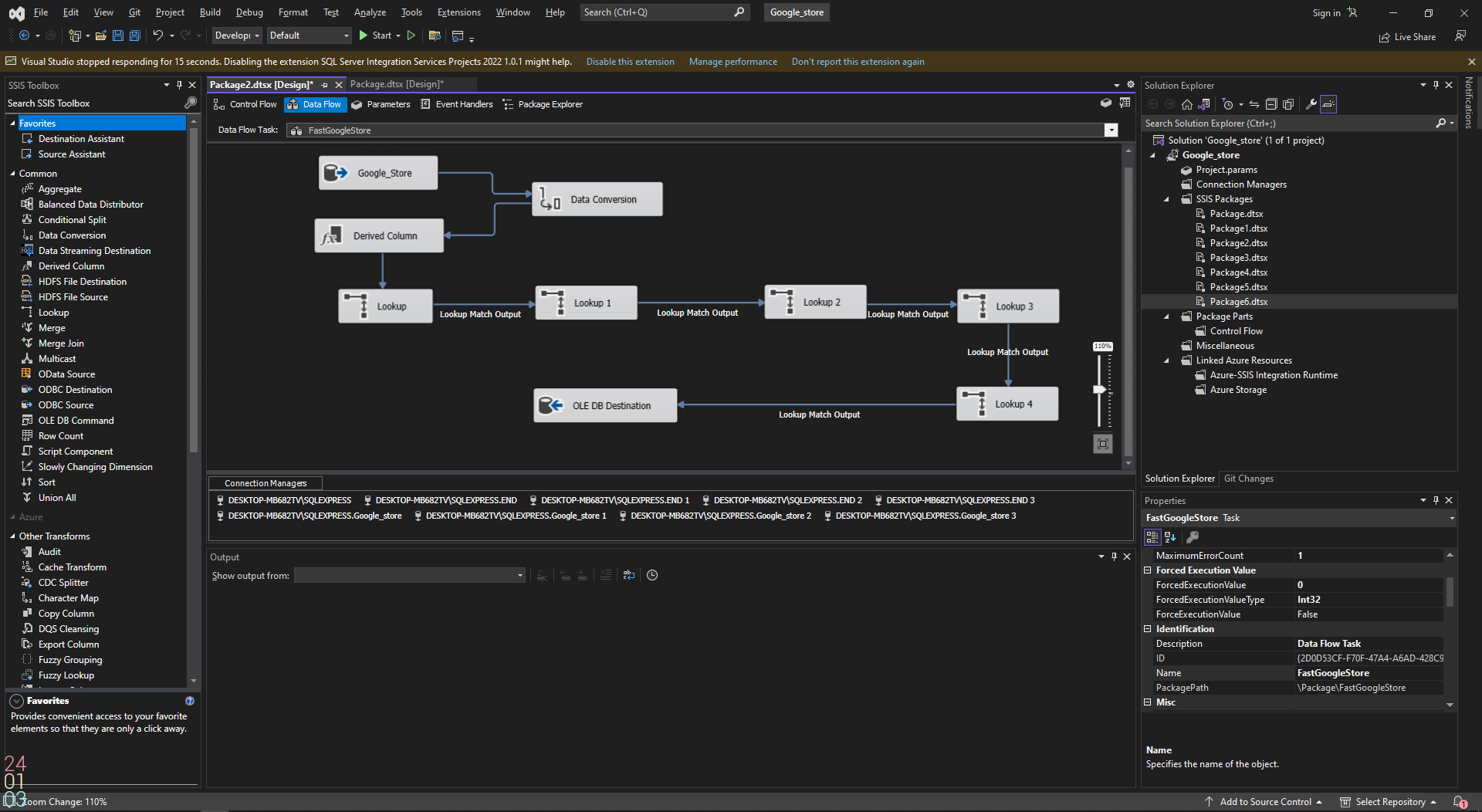


Bước 5:

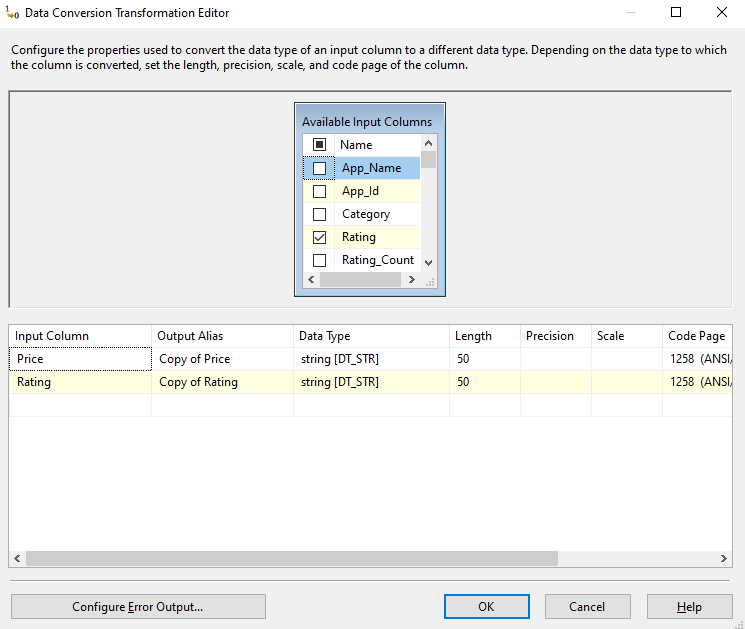
Tạo bảng FastGoogleStore, đầu tiên tạo dataflow task



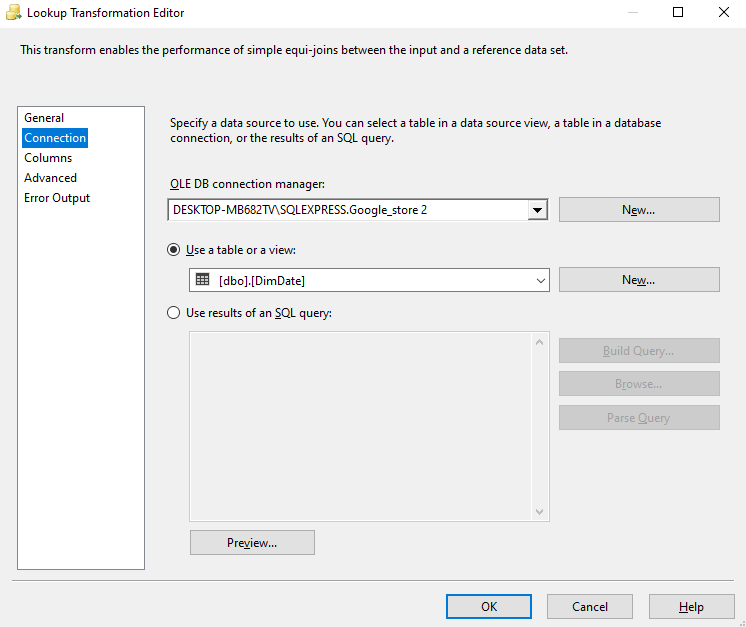
Tạo các khối lệch trong dataflow task như sau



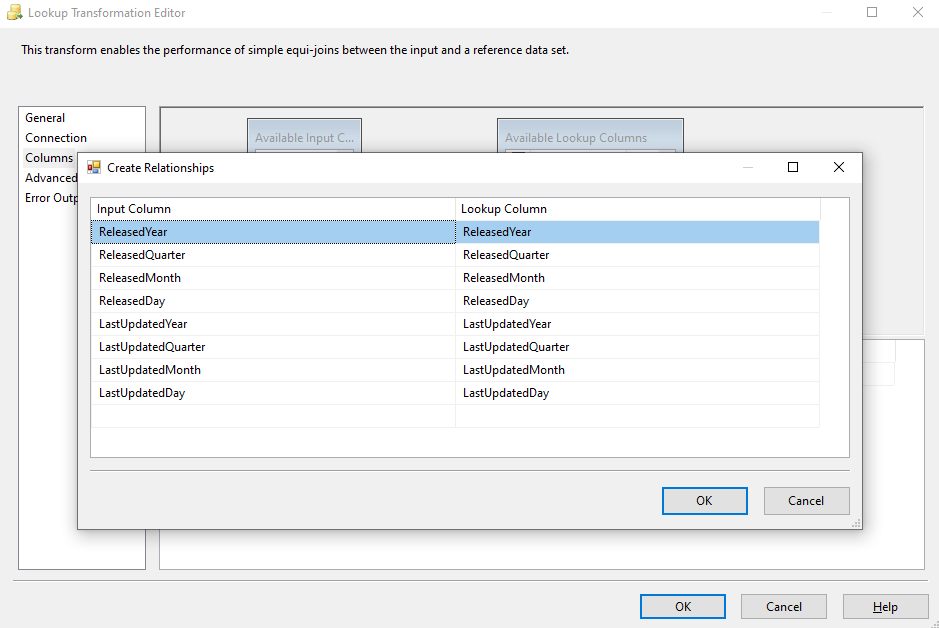
Lưu ý cần conversion cột Price và Rating đễ về data type string trước khi tiến hành lookup



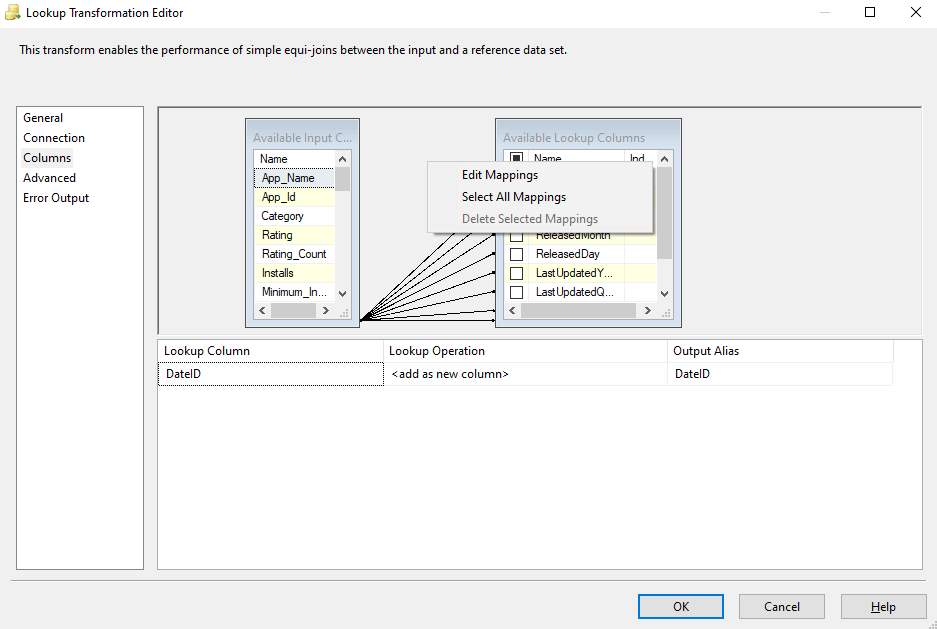
Tạo các toolbox lookup sử dụng các bảng Dim vừa tạo đễ tạo ra bảng Fast, tại tab Connection chọn bảng Dim phù hợp.



Tại tab Columns click chuột phải vào khoãng không trong vùng Mappings sao đó chọn Edit Mappings và lựa chọn các Input Column và Lookup Column thích hợp

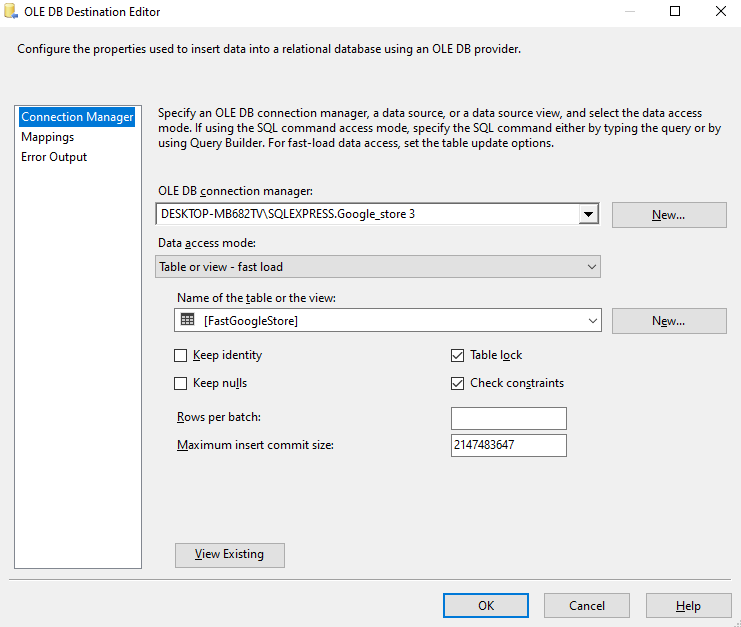


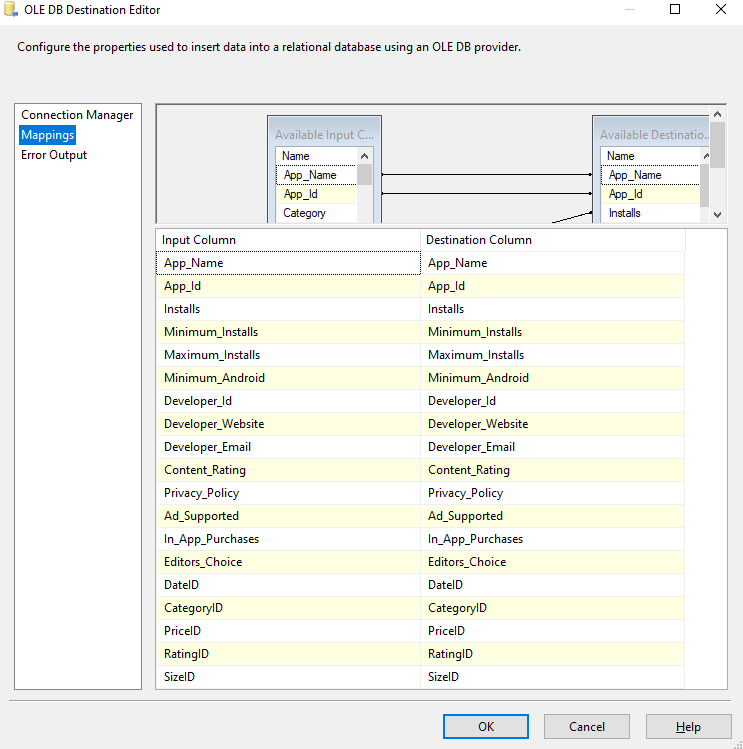
Tạo thêm cột thích hợp



Tiếp theo destination vào SQL và tạo bảng Fast thích hợp

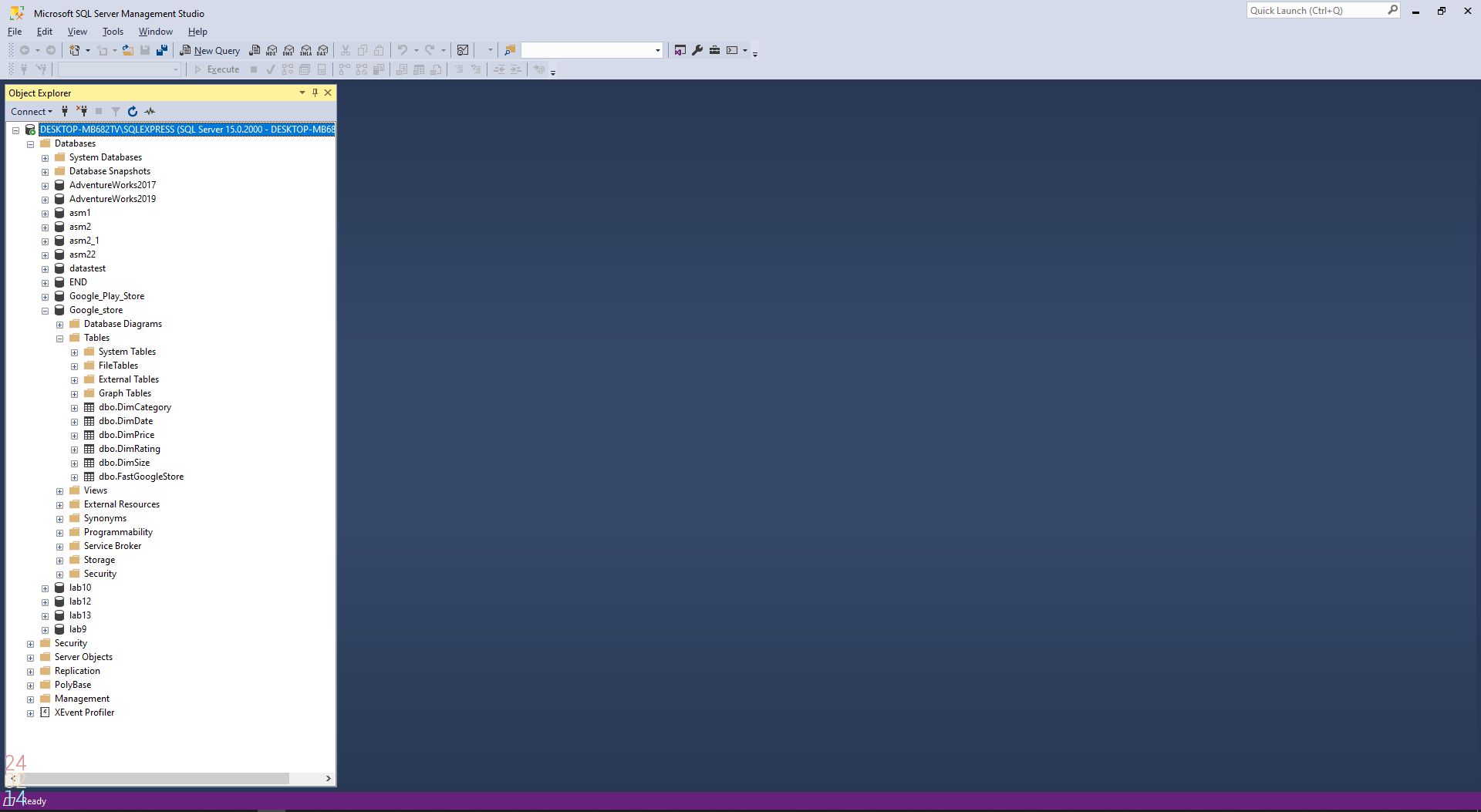
Lưu ý loại bỏ các cột không cần thiết





Bước 6:

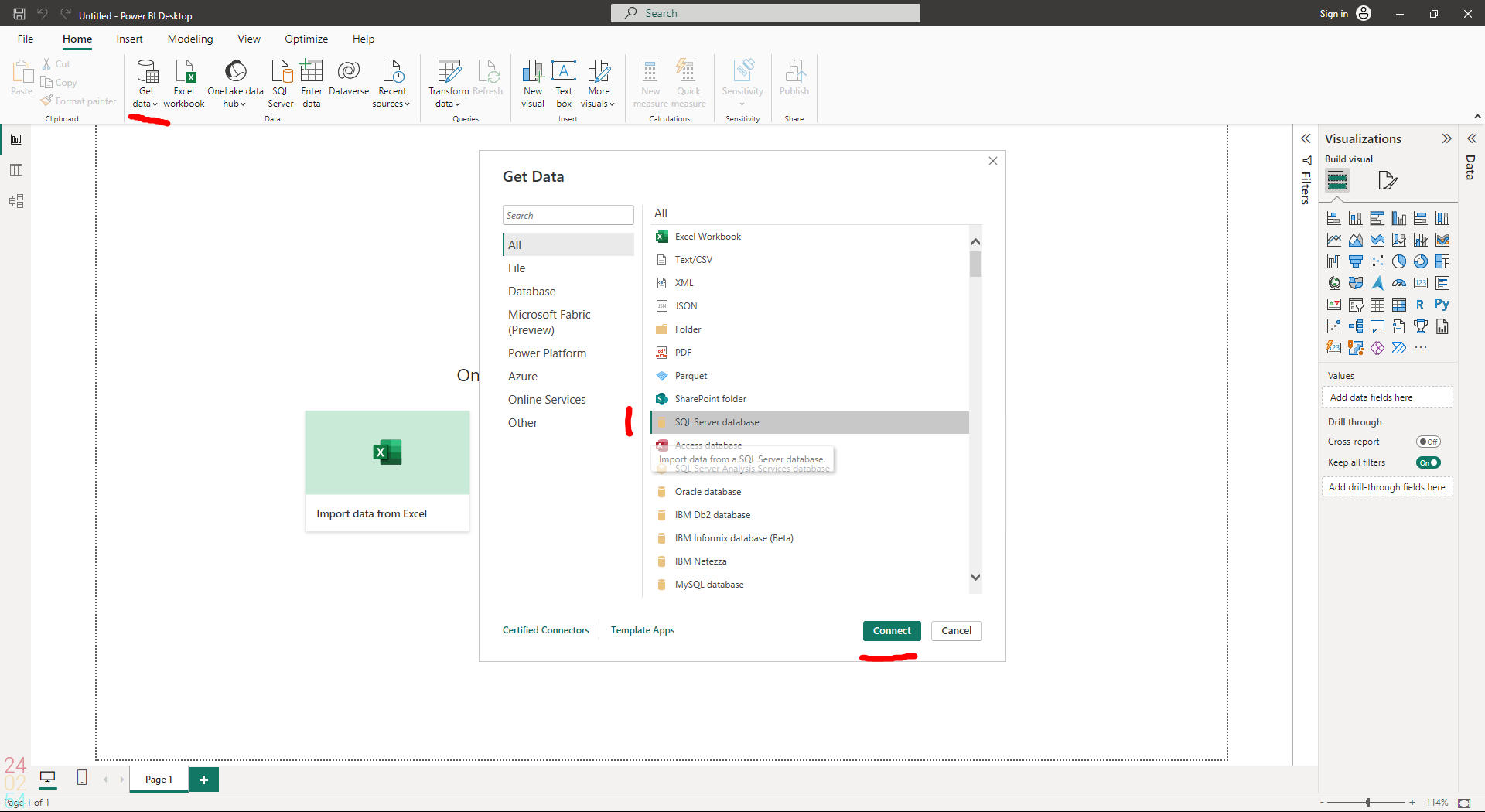
Kiếm tra các bảng Dim và Fast đã xây dựng với SQL server



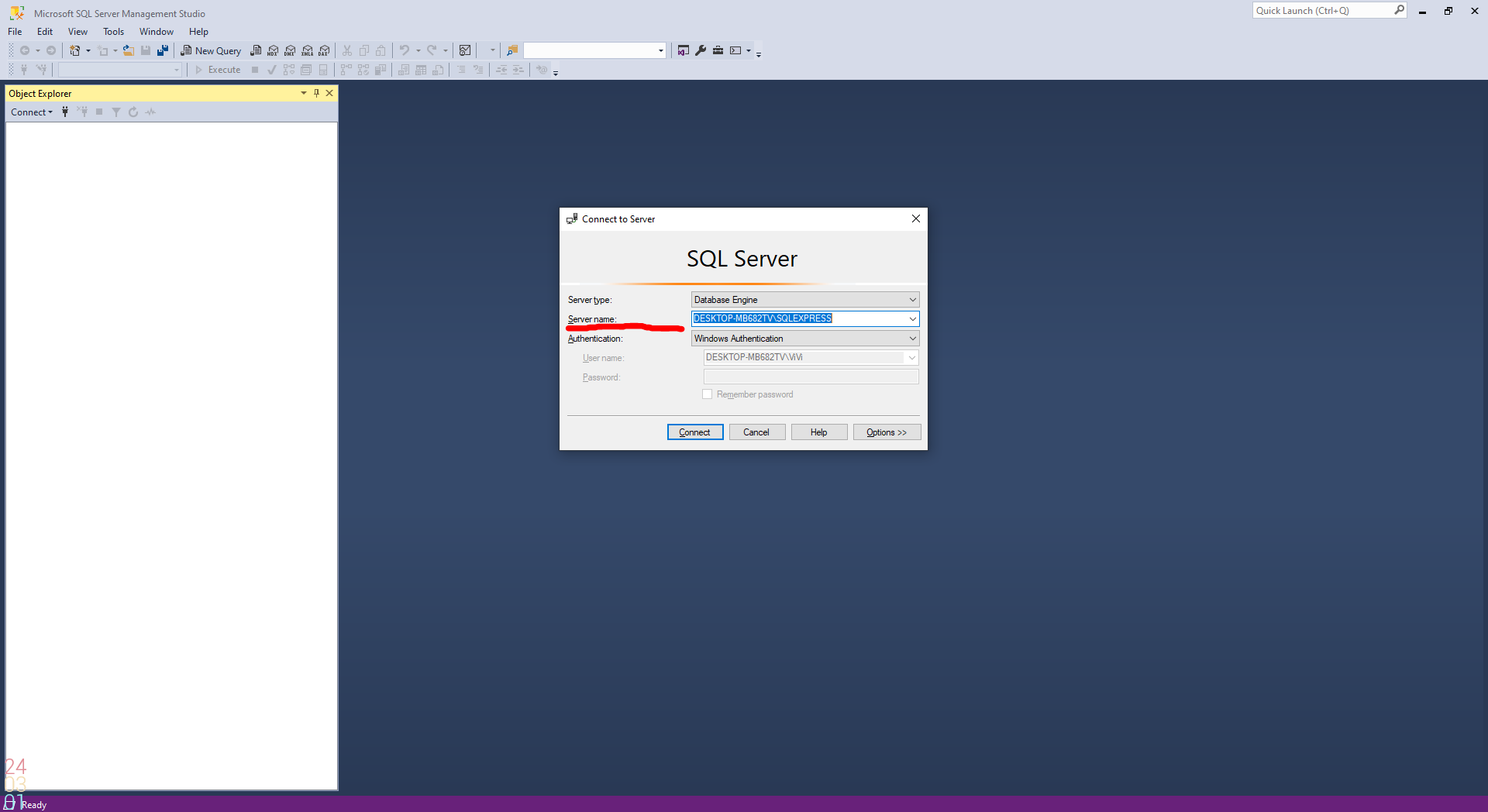
Bước 7:

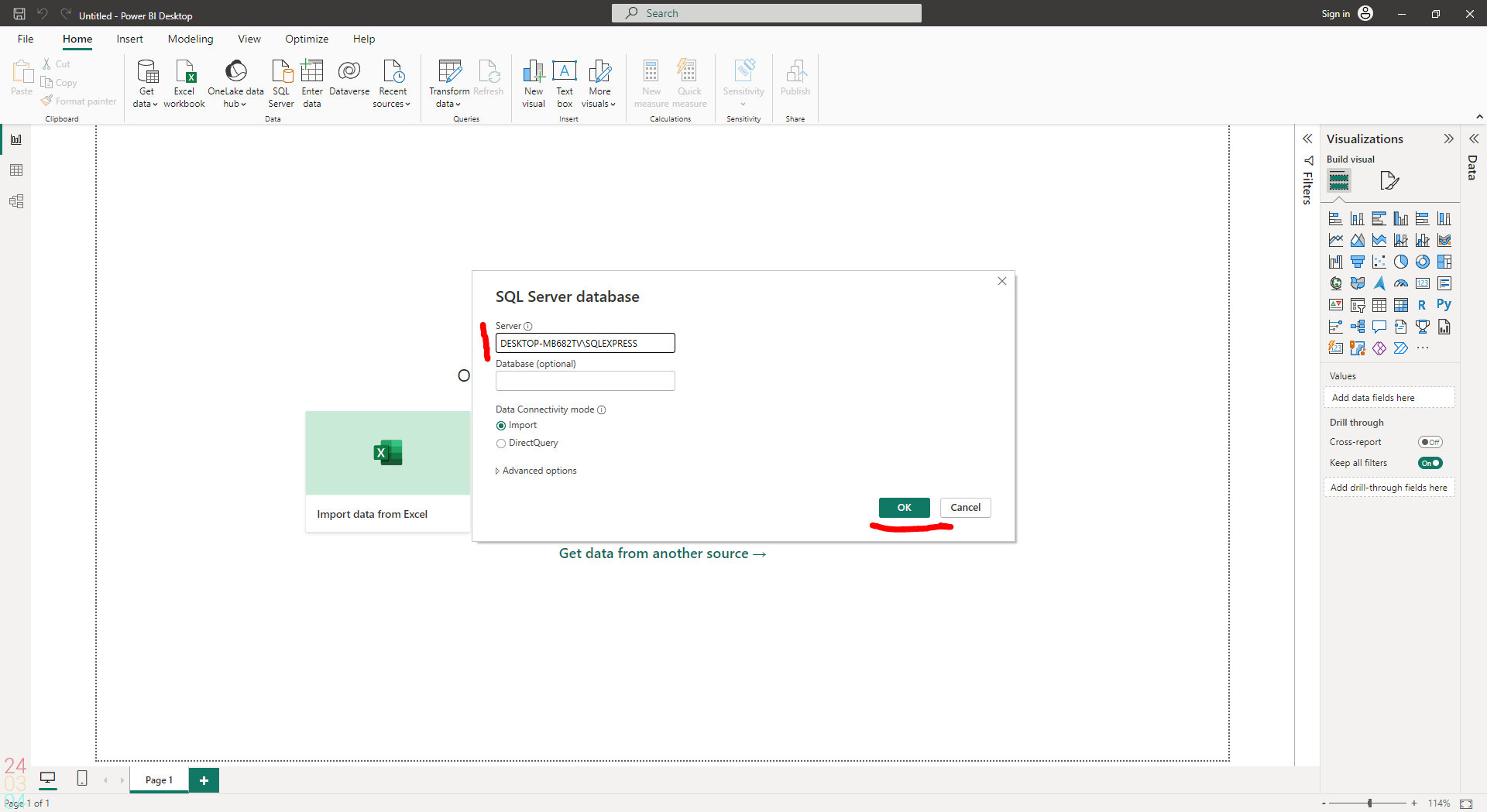
Thực hiện tải dữ liệu vào Power BI

Mỡ Power BI desktop và chọn Get data thông qua SQL server database

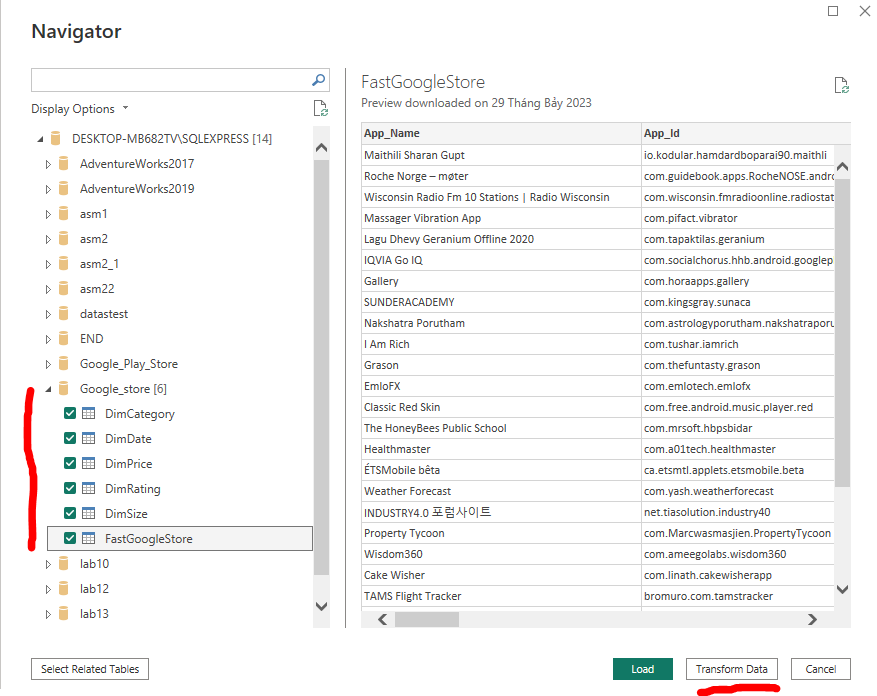


Điền server name DESKTOP-MB682TV\SQLEXPRESS

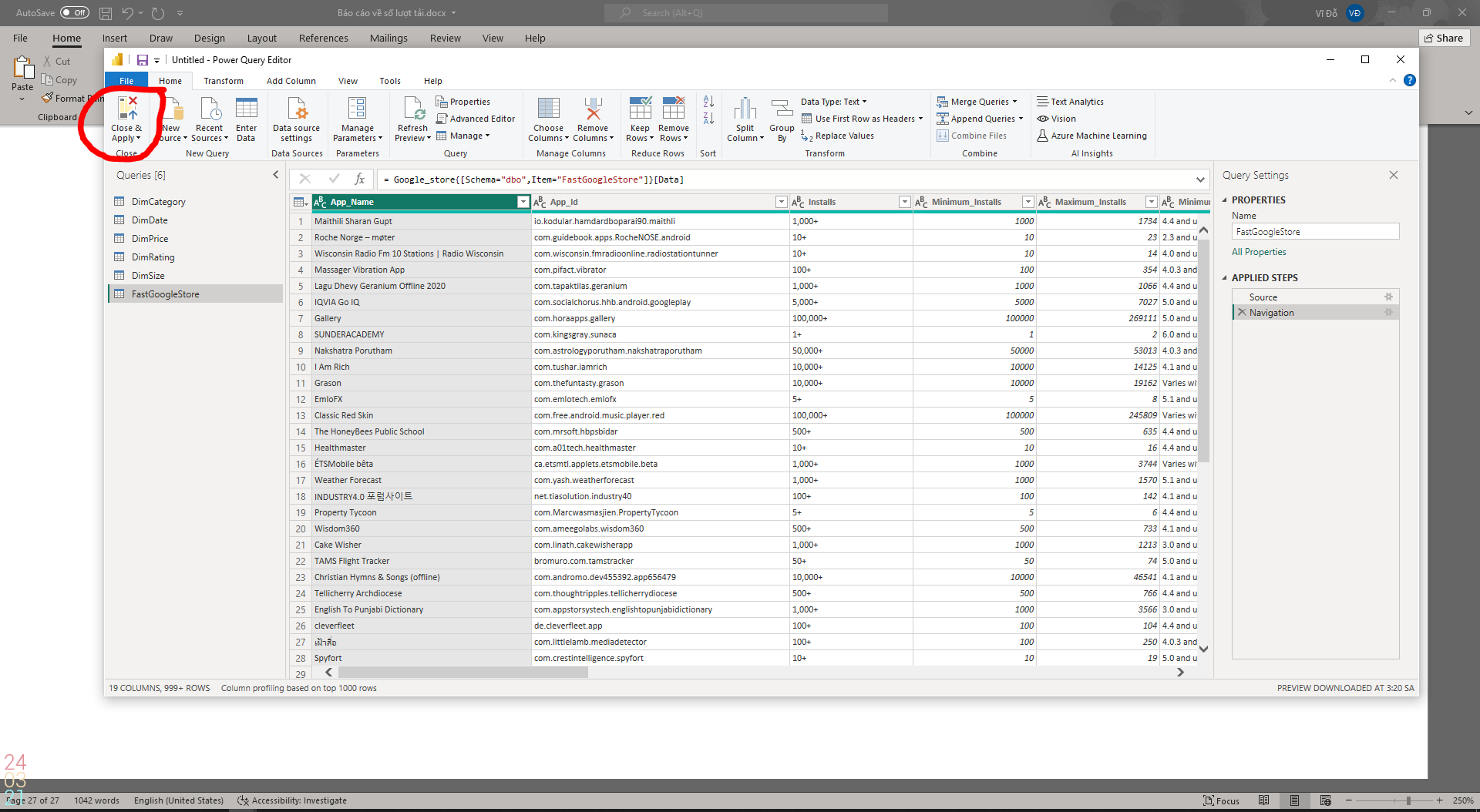




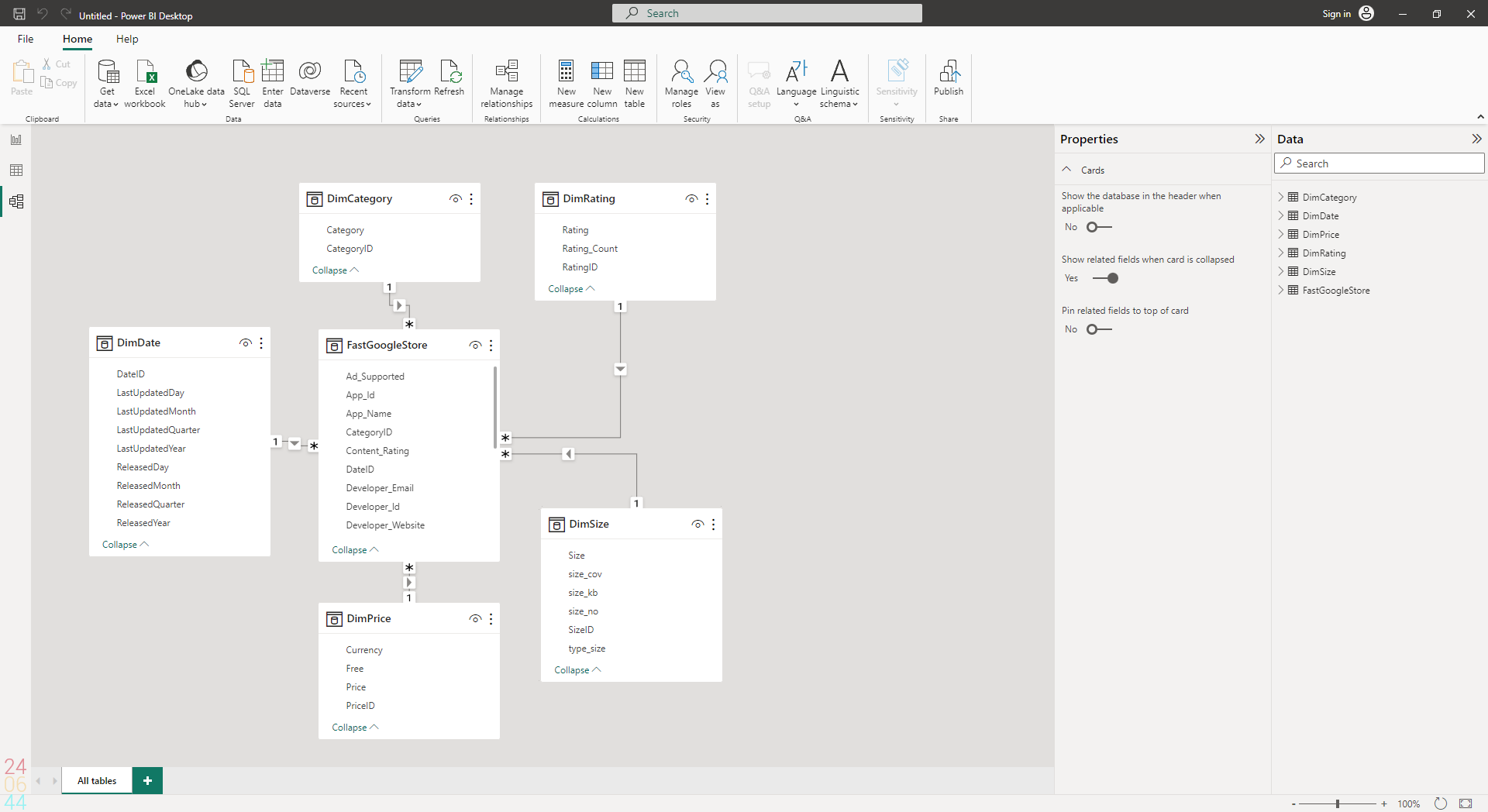
Lựa chọn database thích hợp và chú ý transform data phù hợp



Sau khi kiểm tra dư liệu đầu vào phù hợp thì nhấn Close & Apply



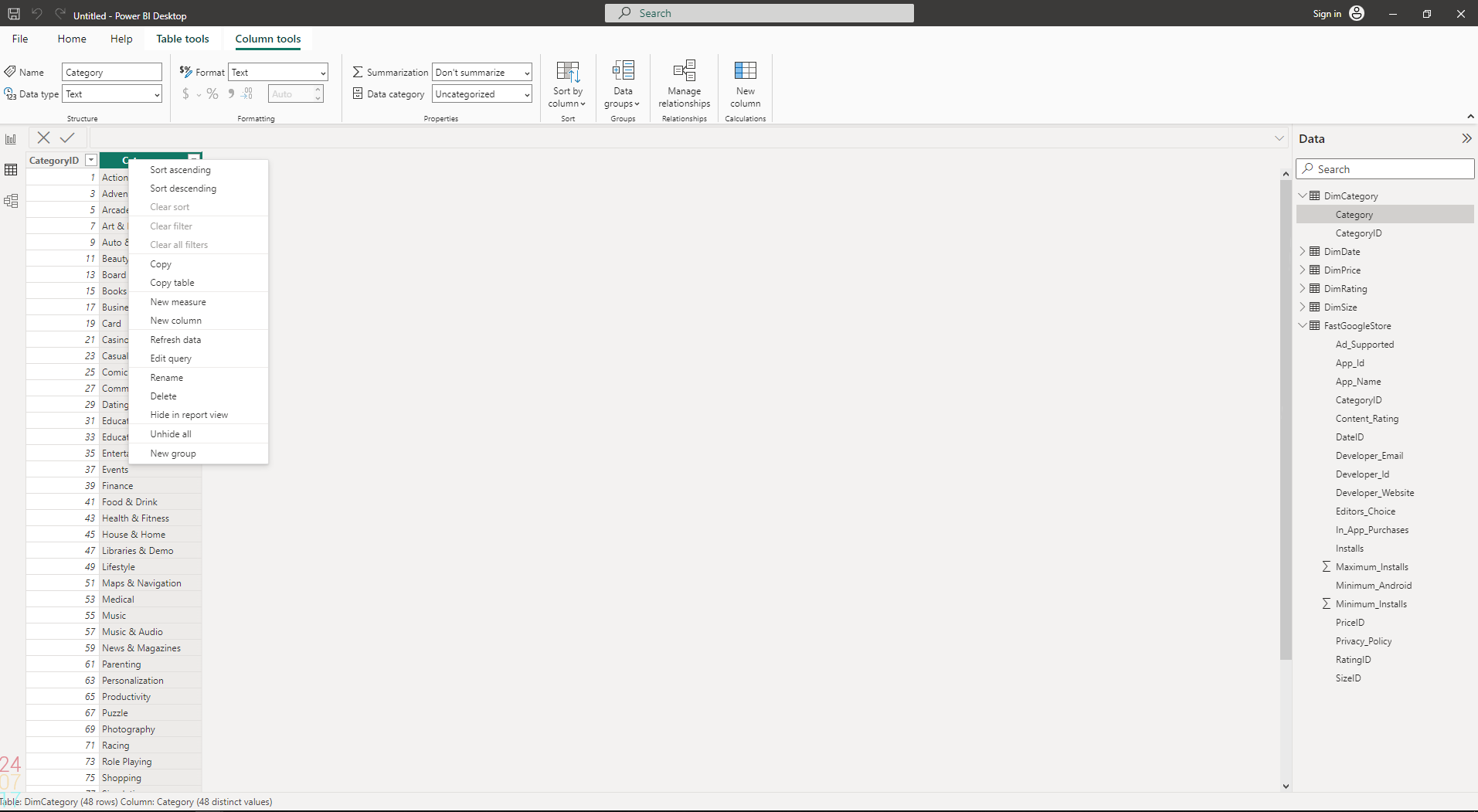
Sau khi tải dữ liêu lên Power BI thu được các table và các relationship như hình



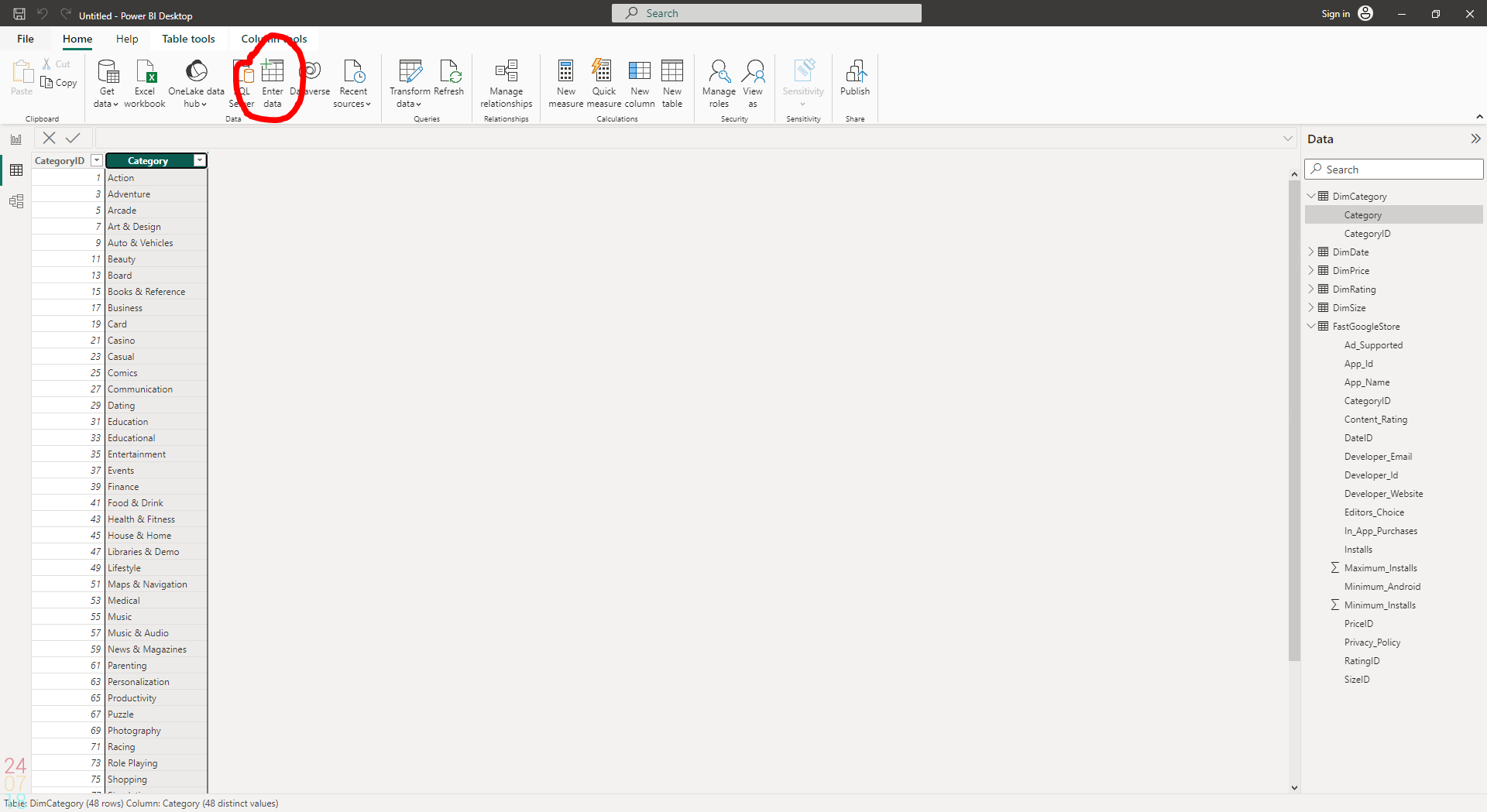
Bước 8:

Tạo thêm các table mới đễ dễ dàng thực hiện visual live

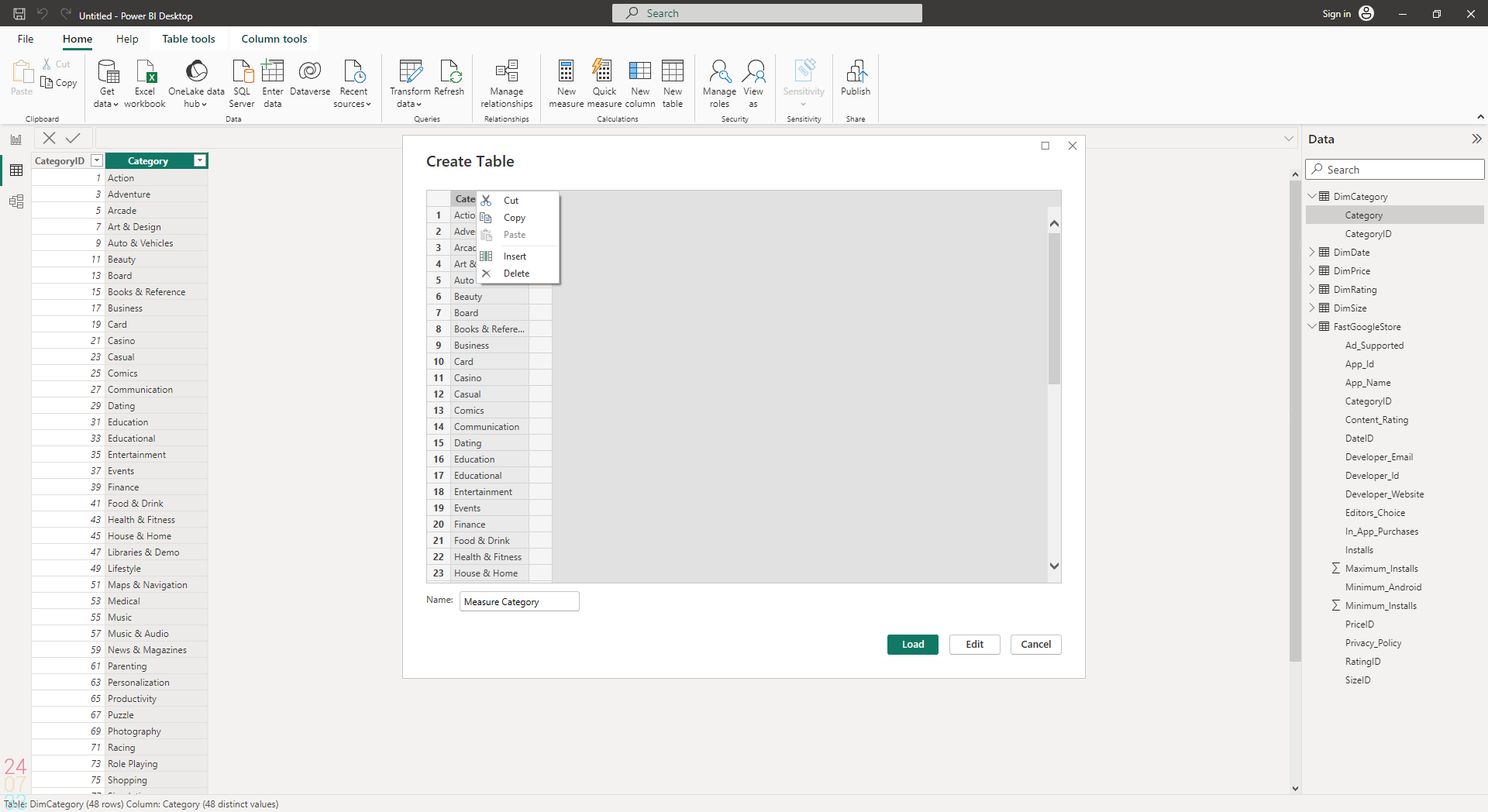
Tại tab Table view và bảng DimCategory, click chuột phải vào cột Category và chọn Copy



Tại tab Home và chọn Enter data

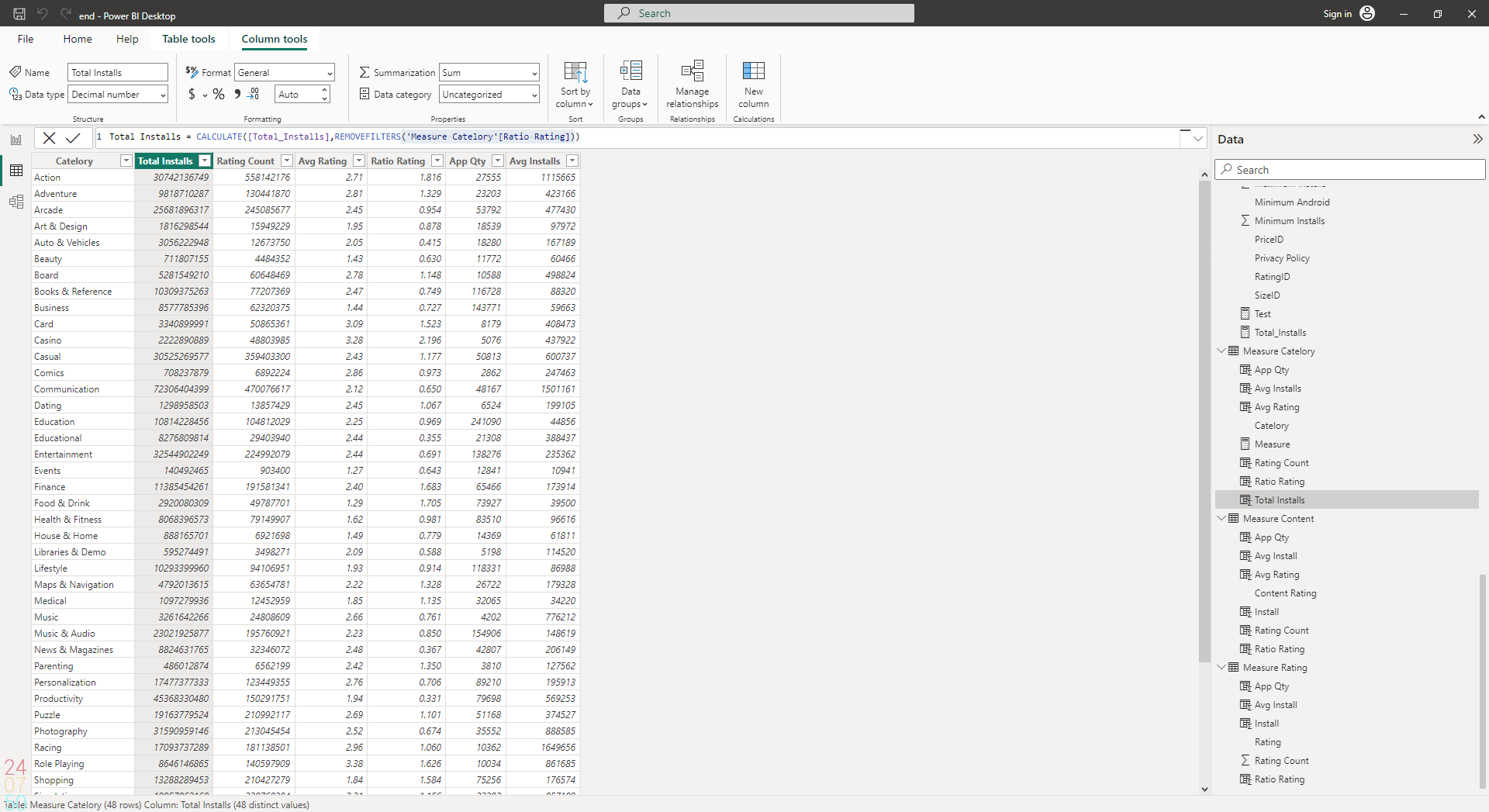


Click chuột phải và chọn Paste, sau đó đặt lại tên cho tên table Measure Category



Tạo cột Total Installs với công thức Total Installs = CALCULATE ([Total\_Installs],REMOVEFILTERS('Measure Catelory'[Ratio Rating])) trong đó measure Total\_Installs = SUM (GooglePlayStoreFact[Maximum Installs] )

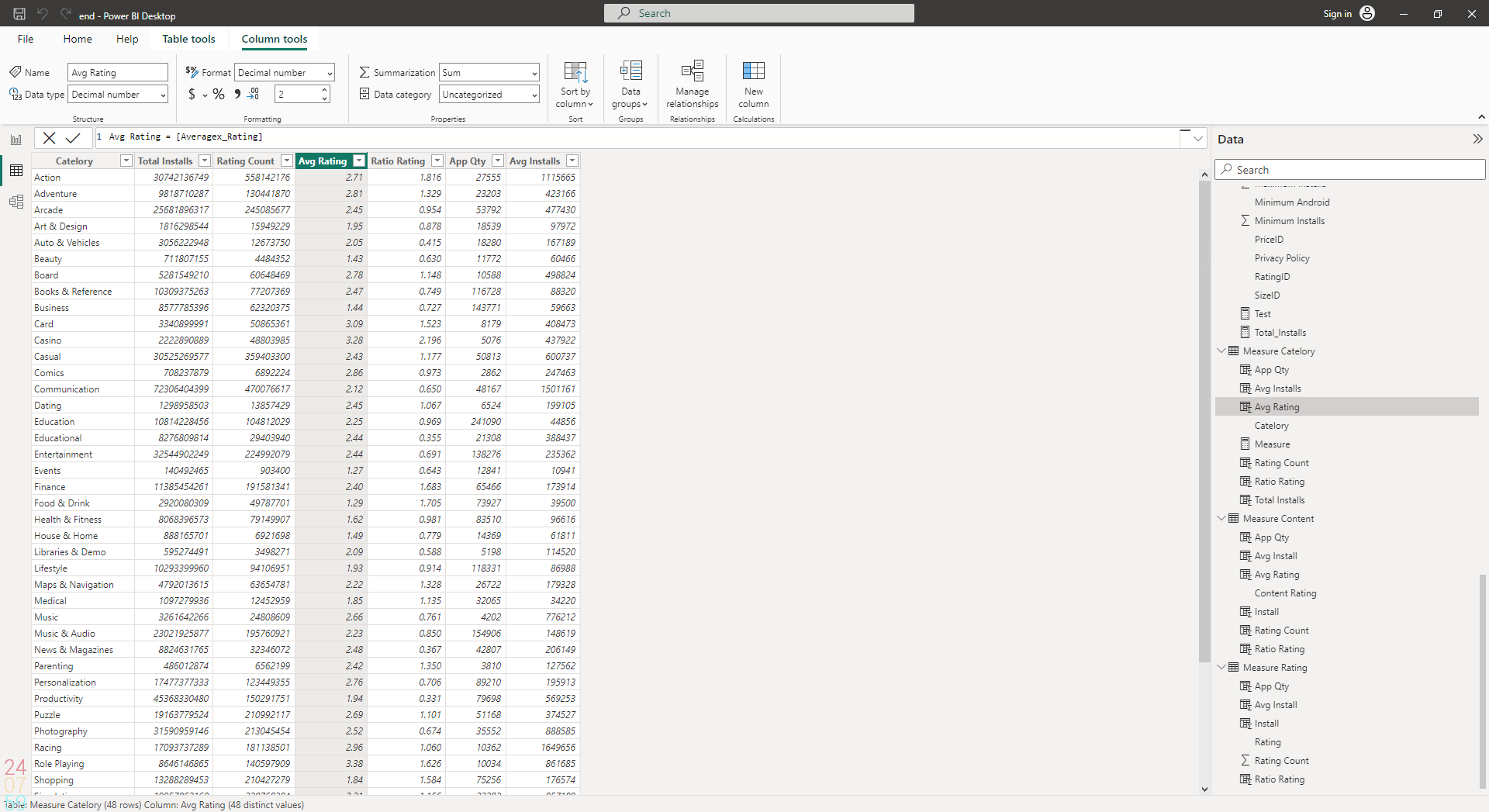
Lưu ý REMOVEFILTERS('Measure Catelory'[Ratio Rating]) dùng đễ loại bỏ bộ lọc đễ có thể thựt thi các các cột khác



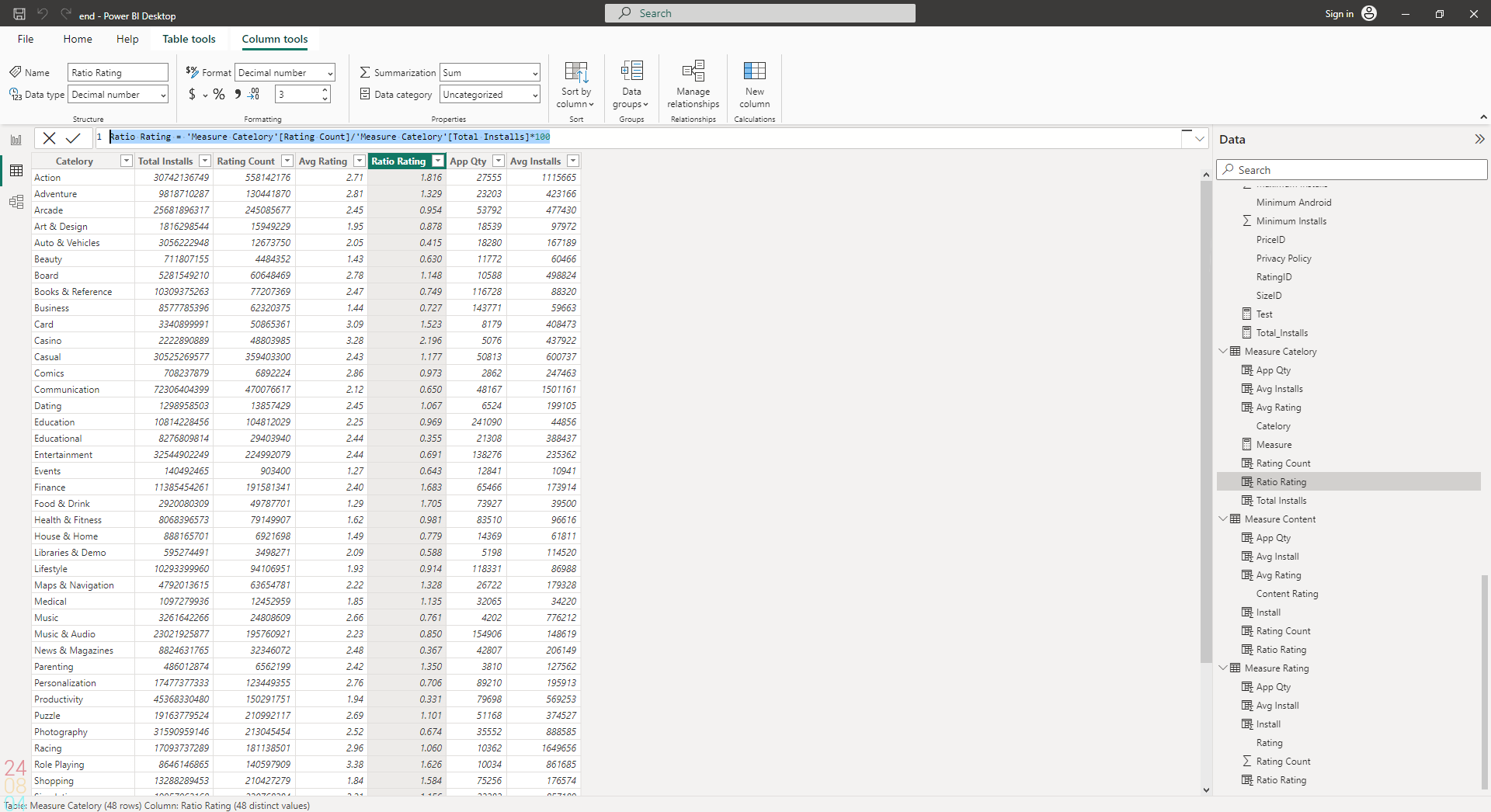
Tạo cột Rating Count với công thức Rating Count = CALCULATE (SUMx(GooglePlayStoreFact,[Sum RatingCount]),REMOVEFILTERS('Measure Catelory'[Ratio Rating])) trong đó measure Sum RatingCount = CALCULATE(sumx(GooglePlayStoreFact,sum(DimRating[Rating Count])))



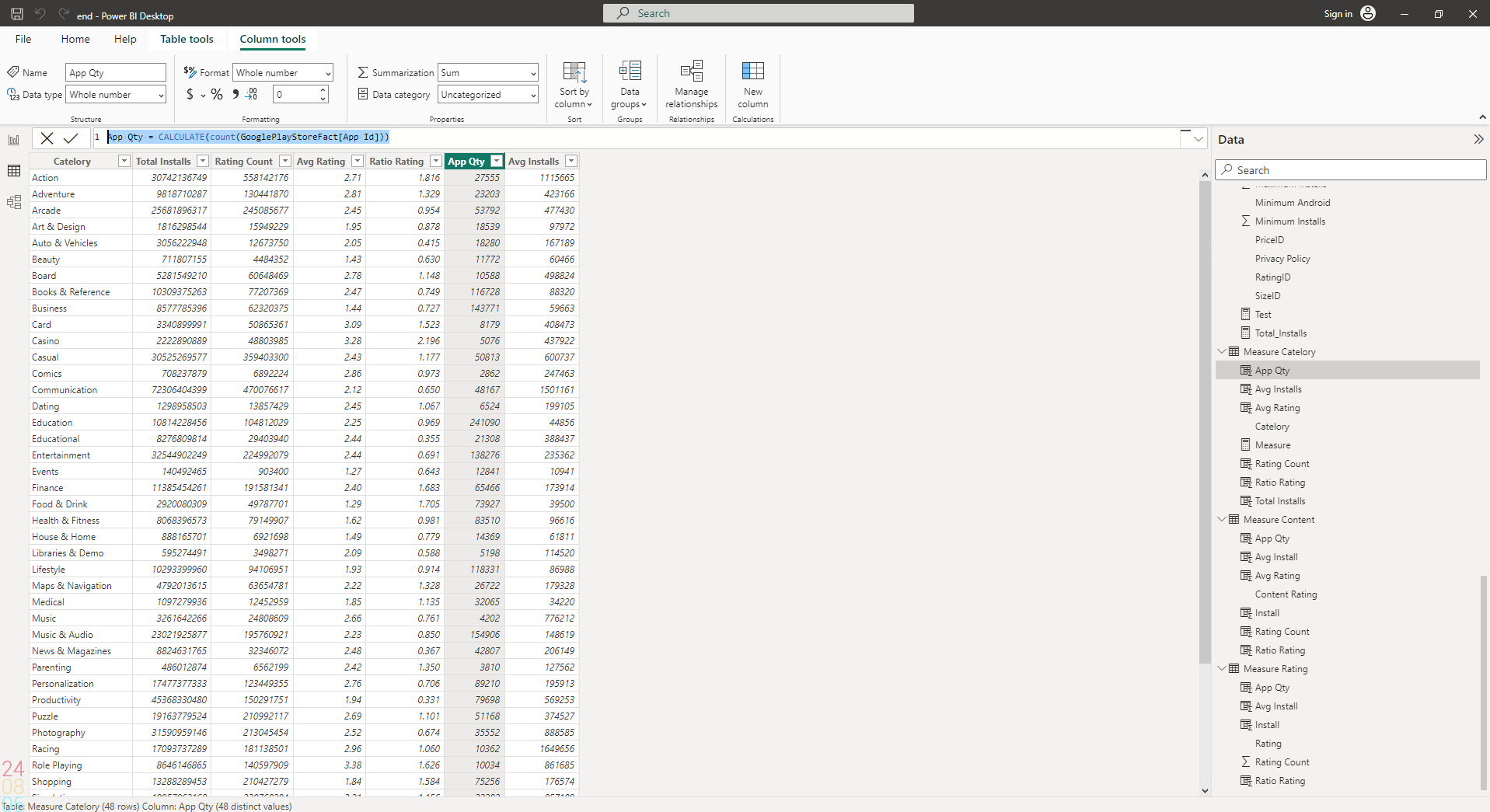
Tạo cột Avg Rating với công thức Avg Rating = [Averagex\_Rating] trong đó measure Averagex\_Rating = AVERAGEX(GooglePlayStoreFact,[Avg\_Rating])



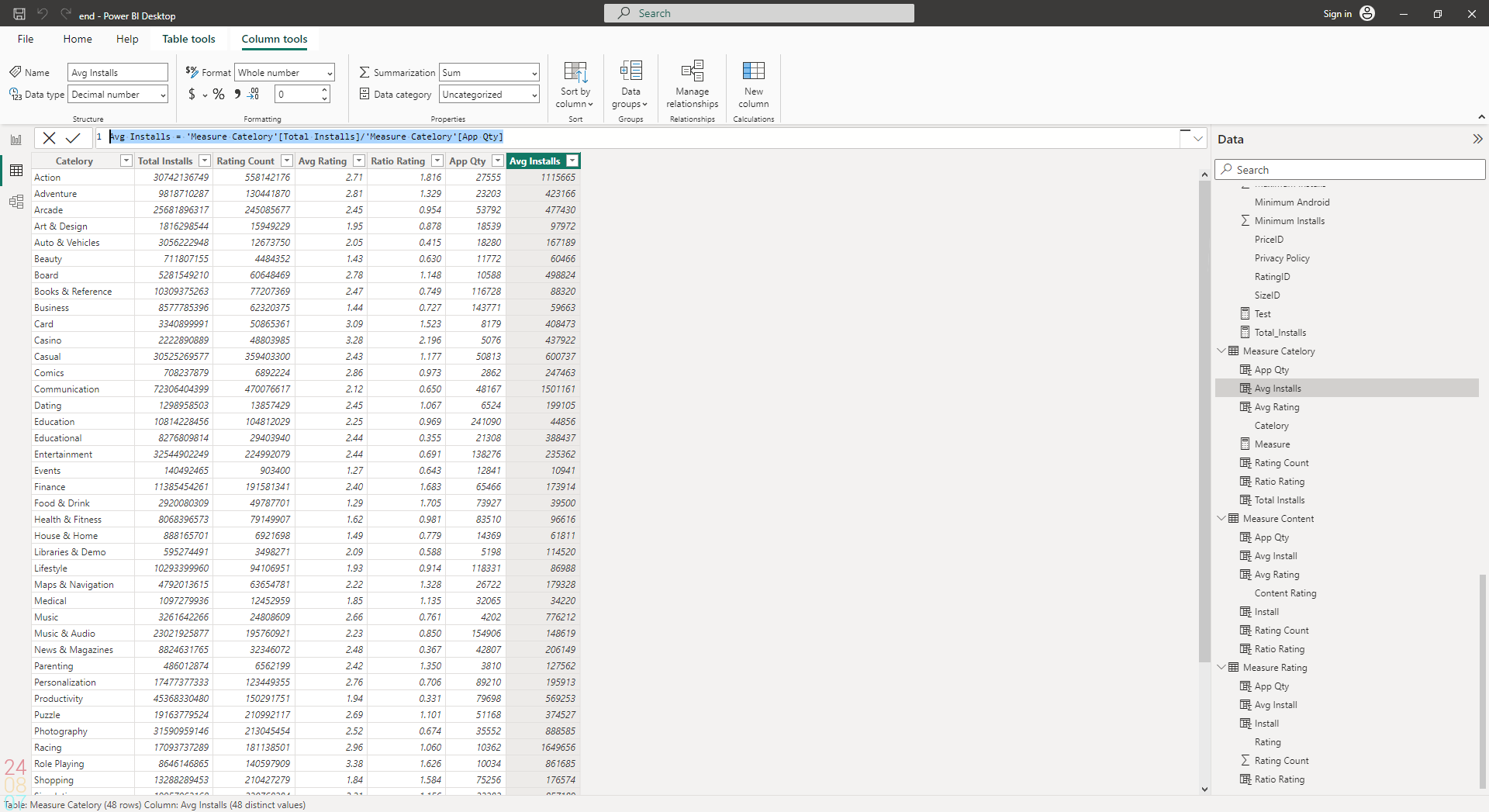
Tạo cột Ratio Rating với công thức Ratio Rating = 'Measure Catelory'[Rating Count]/'Measure Catelory'[Total Installs]\*100



Tạo cột App Qty với công thức App Qty = CALCULATE(count(GooglePlayStoreFact[App Id]))



Tạo cột Avg Installs với công thức Avg Installs = 'Measure Catelory'[Total Installs]/'Measure Catelory'[App Qty]

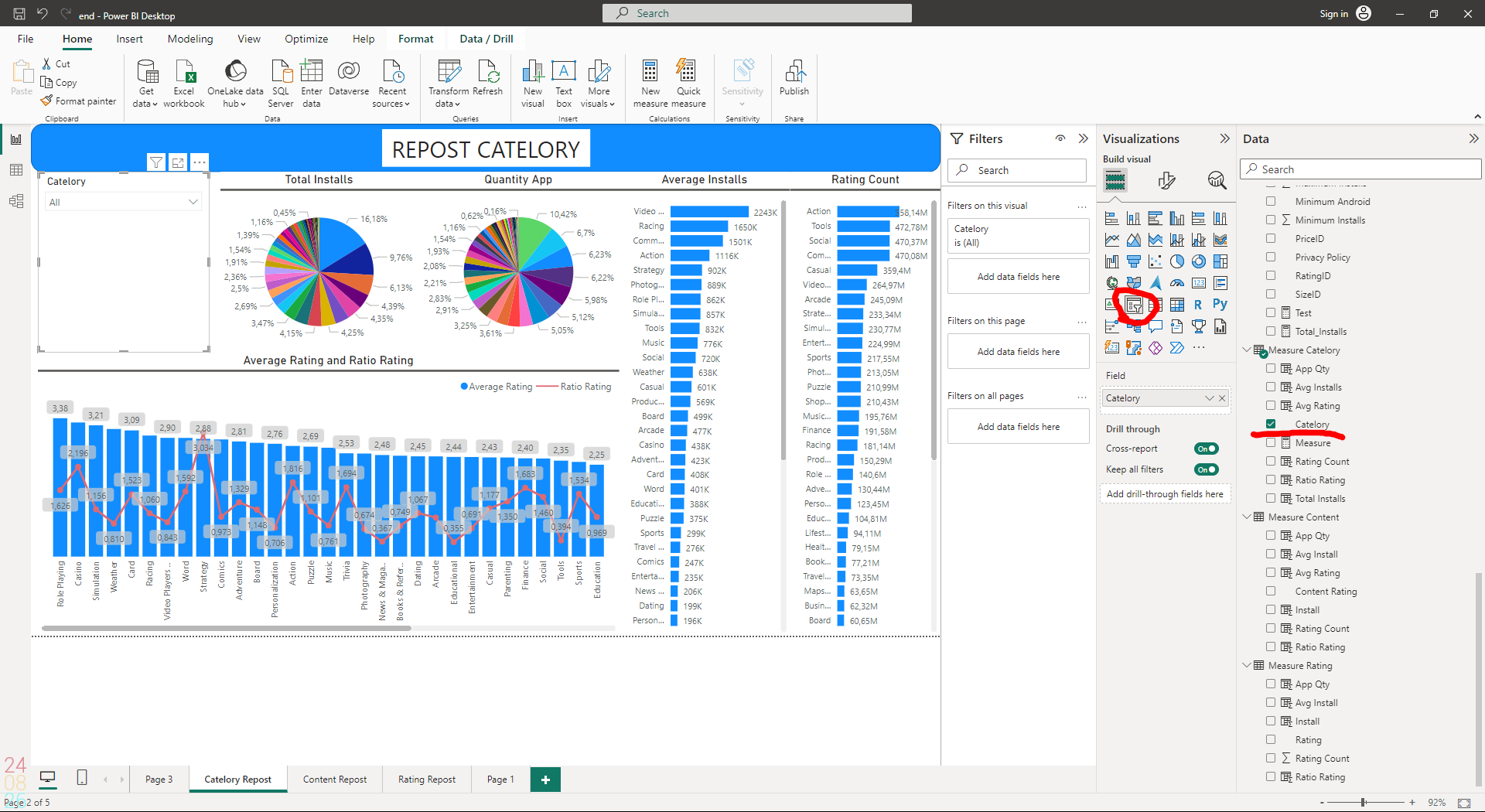


Thực hiện tương tự với bảng Measure Content và Measure Rating

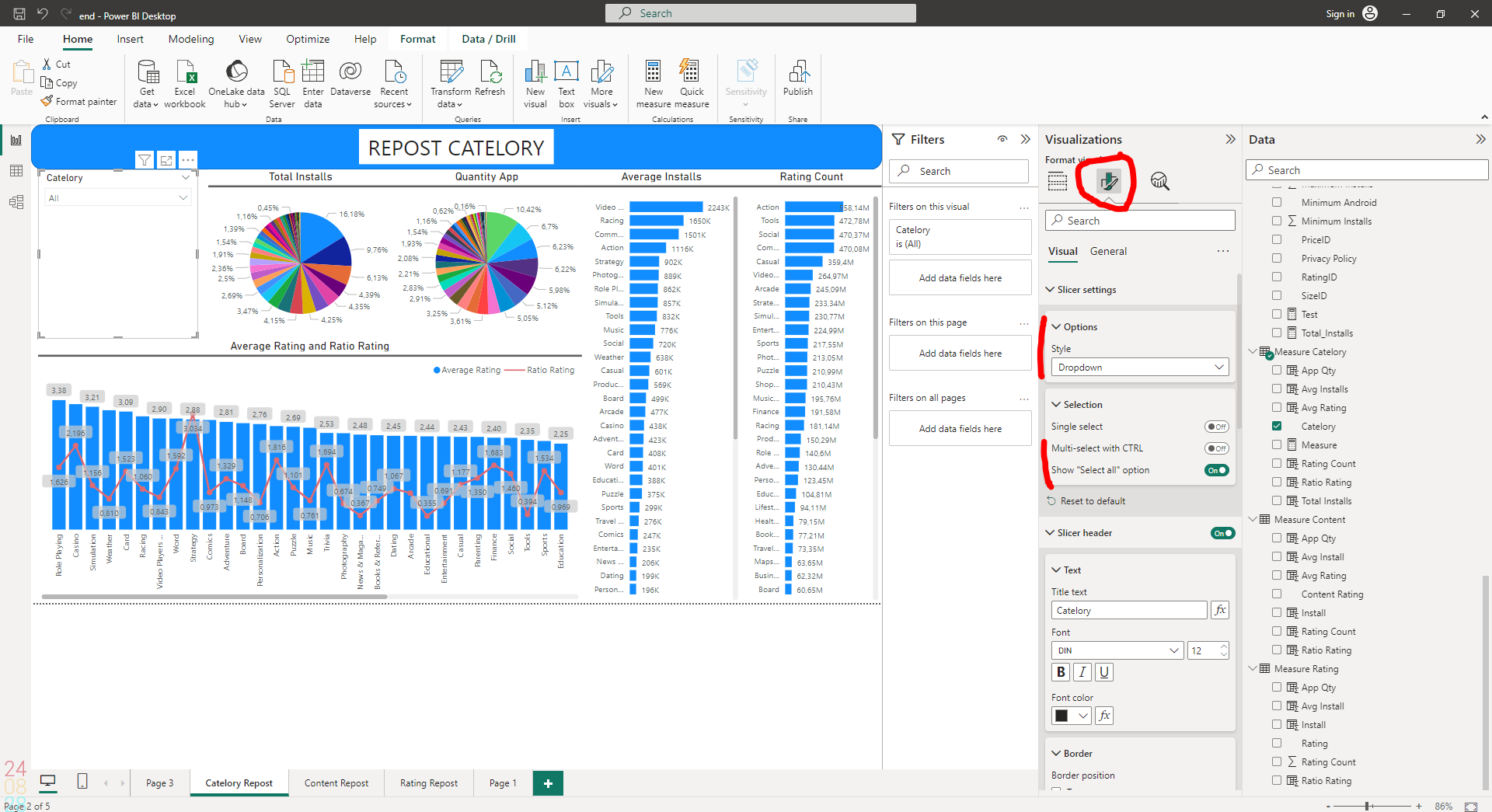
Bước 9:

Thực hiện visual line các bảng vừa tạo

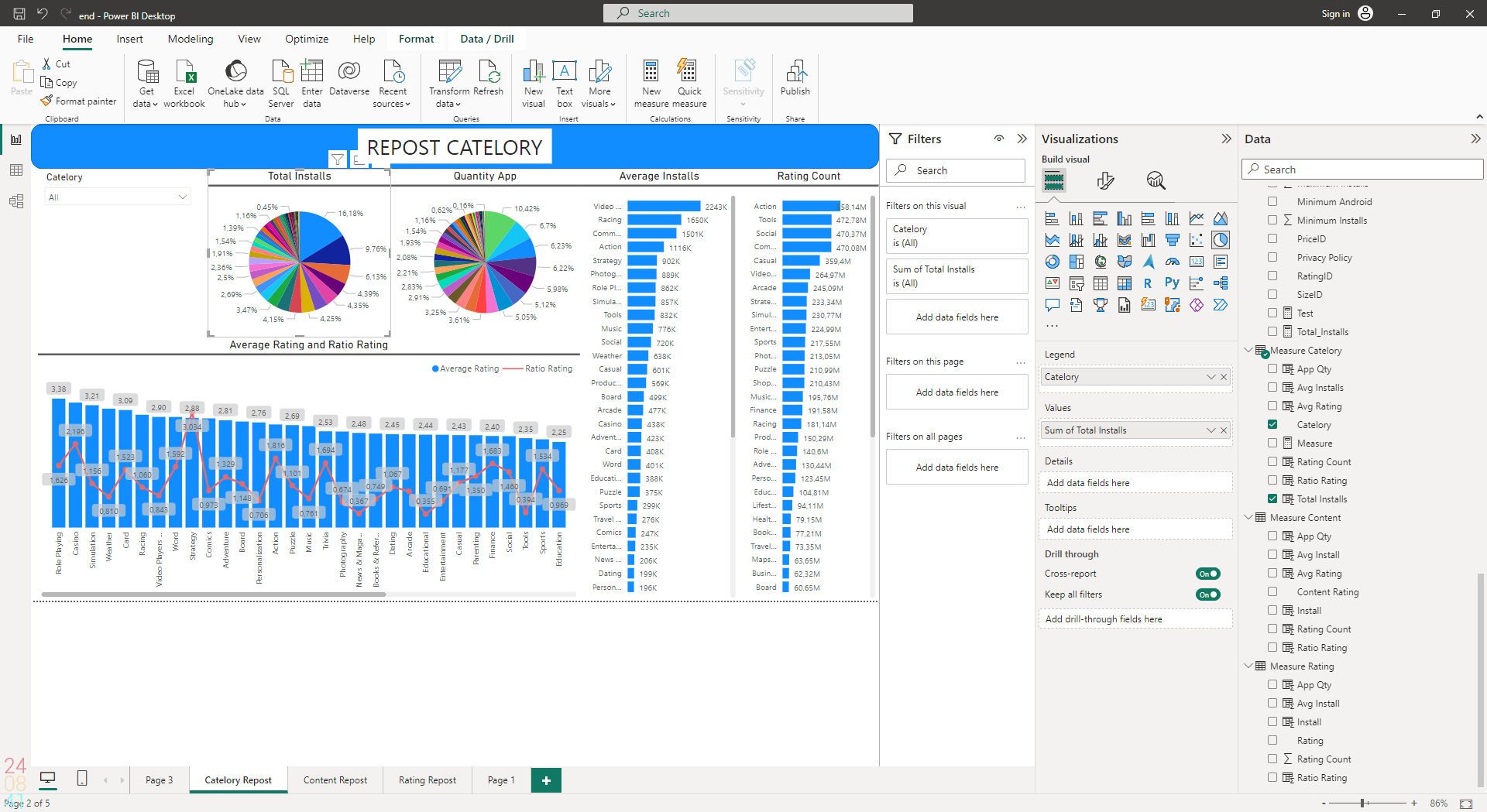
Tại tab Report View tạo Slicer bằng cách click chuột trái vào icon slicer ở phần Visualization và tích chọn Category



Có thể Format visual lại bằng cách click vào icon Format your visual và tuỳ chình phần options



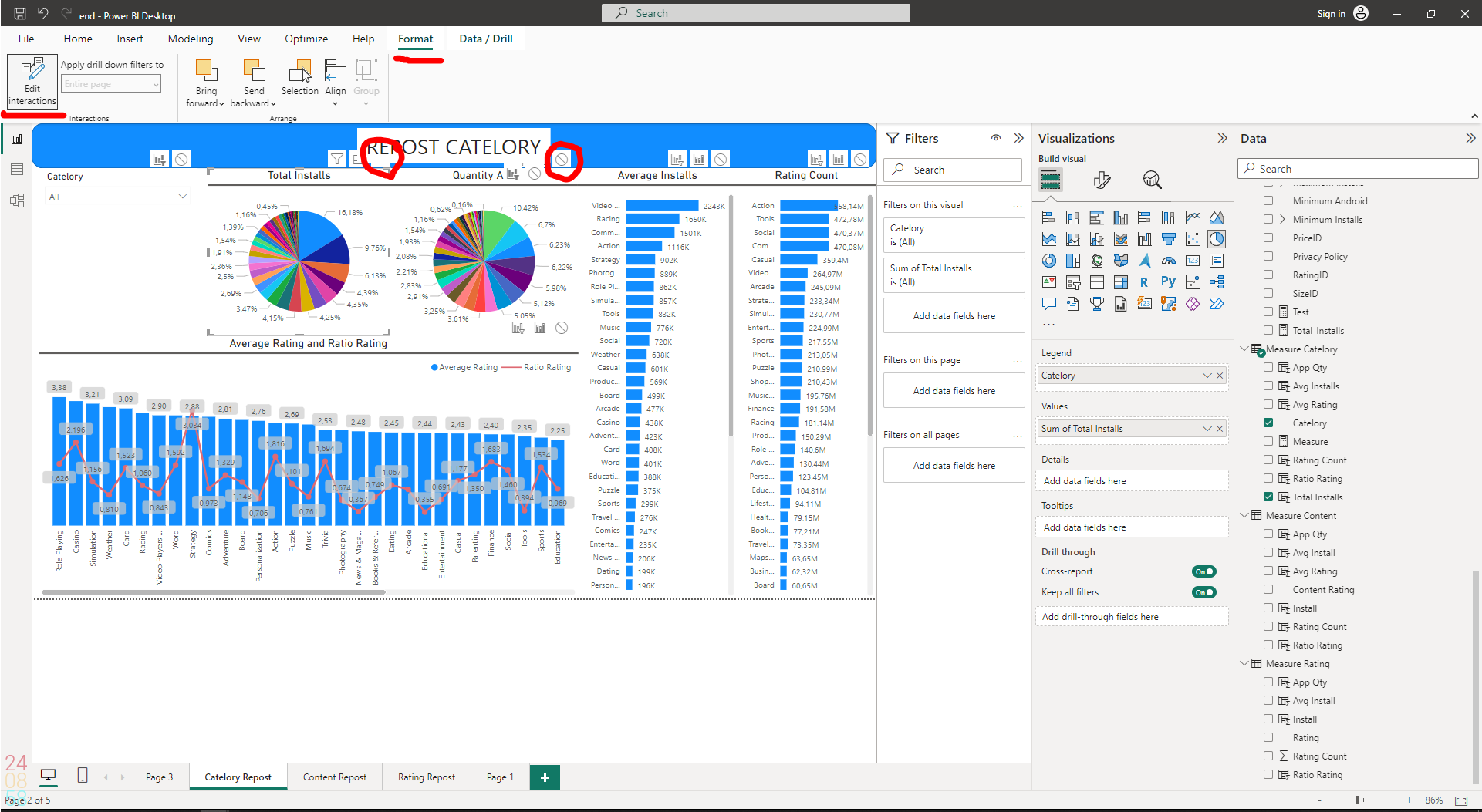
Tạo Pie Chart cho Total Installs bằng cách nhấn vào icon Pie Chart ở phần Visualizations và chọn thông số Category



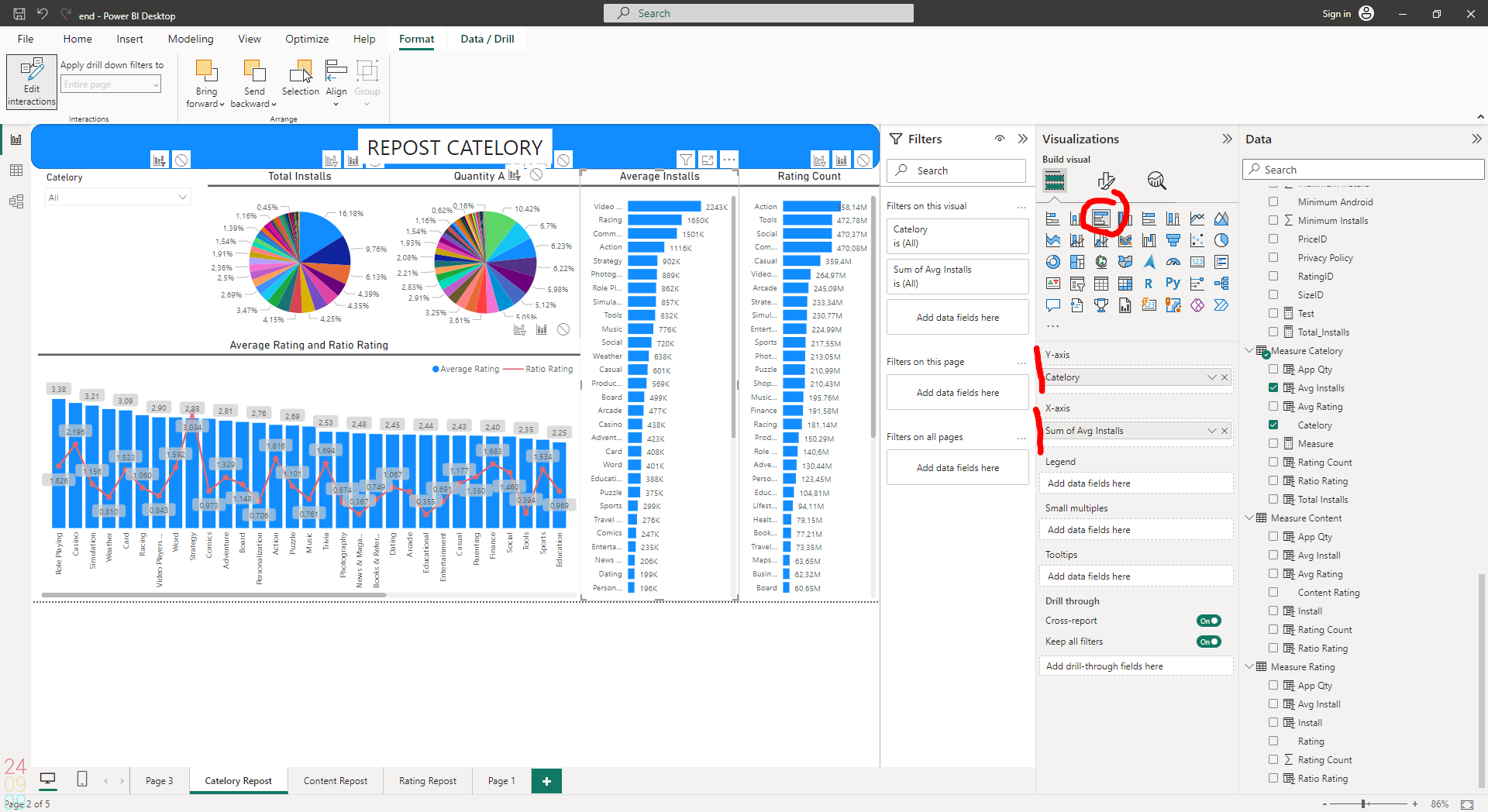
Có thể định dạng pie chart đễ ưu nhìn hơn

Thực hiện tương tự với pie chart của Quantity App

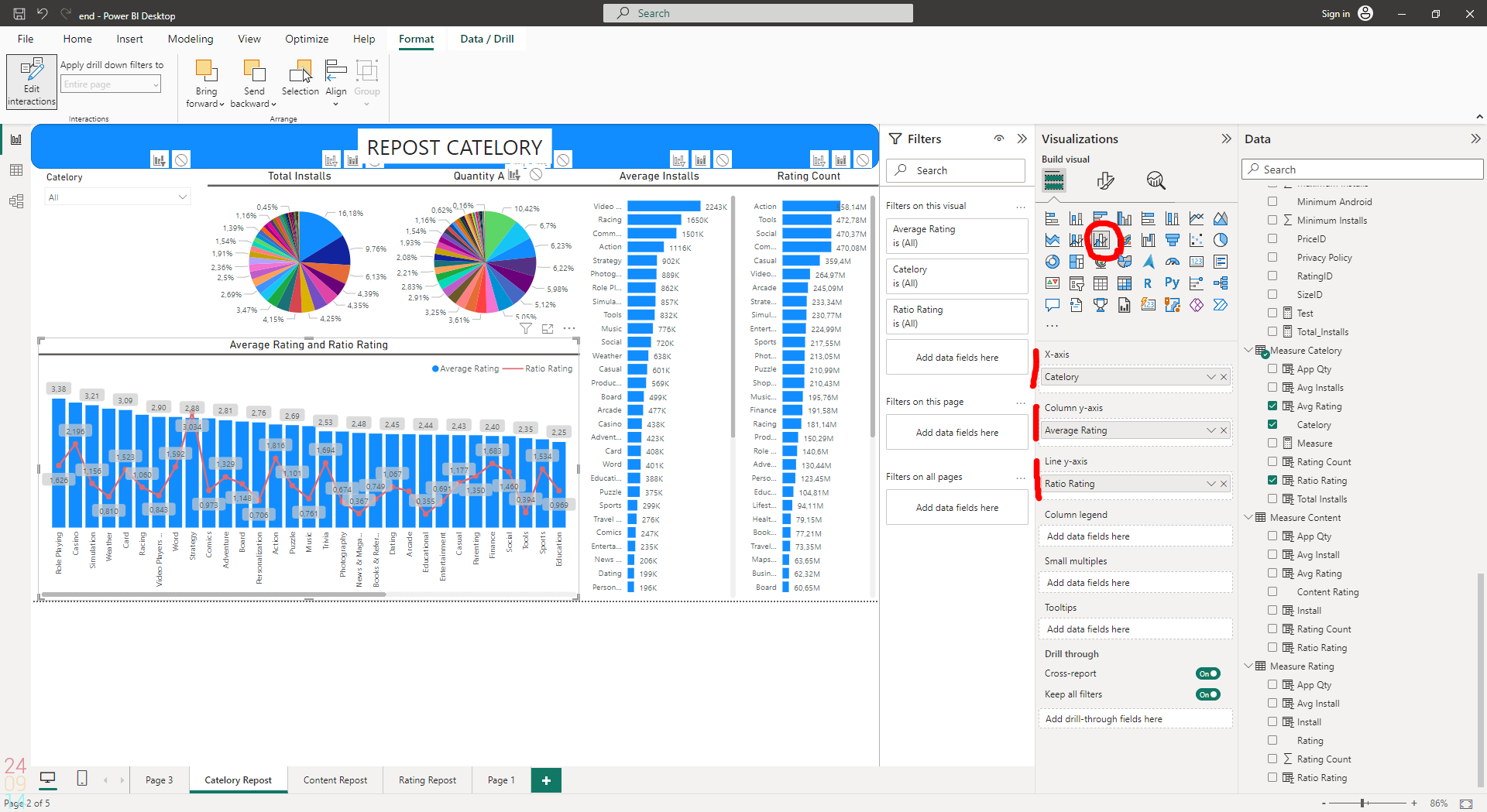
Sao đó edit interactions đễ không link với 2 pie chart này bằng cách click vào Edit interactions sao đó chọn icon None



Tạo bar chart cho Average Installs sử dụng thông số Category và Avg Installs sao đó tuỳ chình format ưa nhìn. Thực hiện tương tự với bar chart Rating Count



Tạo line & column chart Average Rating and Ratio Rating sử dụng thông số Category, Average Rating và Ratio Rating sau đó tuỳ chình format ưa nhìn.



Thực hiện tương tự với các page Content Repost và Rating Repost thu được kết quả như sau

