

# Káº¿t quáº£ hÃ¬nh áº£nh cho logo uit

# LỜI CẢM ƠN

Đầu tiên, nhóm chúng em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới giảng viên ThS. Nguyễn Thị Kim Phụng – người đã giảng dạy và chia sẻ rất nhiều kiến thức cũng như các ví dụ thực tiễn trong các bài giảng. Cô đã hướng dẫn cho chúng em làm bài tập, sửa chữa và đóng góp nhiều ý kiến quý báu giúp chúng em hoàn thành tốt báo cáo môn học của mình.

Bộ môn Kho dữ liệu và OLAP là môn học thú vị, vô cùng bổ ích và có tính thực tế cao. Tuy nhiên, do vốn kiến thức chuyên môn còn nhiều hạn chế và khả năng tiếp thu thực tế còn nhiều bỡ ngỡ. Mặc dù chúng em đã cố gắng hết sức nhưng chắc chắn bài báo cáo khó có thể tránh khỏi những thiếu sót và nhiều chỗ còn chưa chính xác, chúng em rất mong nhận được sự góp ý, chỉ bảo thêm của cô nhằm hoàn thiện những kiến thức của mình để nhóm chúng em có thể dùng làm hành trang thực hiện tiếp các đề tài khác trong tương lai cũng như là trong học tập và làm việc sau này.

Một lần nữa, nhóm xin gửi đến cô, bạn bè lời cảm ơn đặc biệt chân thành và

tốt đẹp nhất!

Thành phố Hồ Chí Minh, tháng 7 năm 2024

**Nhóm sinh viên thực hiện**

Hoàng Ngô Thảo Nguyên

Nguyễn Anh Dĩ

# NHẬN XÉT CỦA GIẢNG VIÊN

……………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….……………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….……………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………………………………………………………………………………………………………………………….………………………

# MỤC LỤC

[LỜI CẢM ƠN i](#_heading=h.1fob9te)

[NHẬN XÉT CỦA GIẢNG VIÊN ii](#_heading=h.3znysh7)

[MỤC LỤC iii](#_heading=h.2et92p0)

[DANH MỤC BẢNG vi](#_heading=h.1t3h5sf)

[DANH MỤC HÌNH ẢNH vii](#_heading=h.4d34og8)

[Chương 1. GIỚI THIỆU TỔNG QUAN 1](#_heading=h.2s8eyo1)

[1. Mô tả bộ dữ liệu 1](#_heading=h.17dp8vu)

[1.1. Mô tả bộ dữ liệu 1](#_heading=h.3rdcrjn)

[1.2. Thuộc tính kho dữ liệu 1](#_heading=h.26in1rg)

[1.3. Mô tả chi tiết thuộc tính 3](#_heading=h.z337ya)

[1.4. Tiền xử lý bộ dữ liệu 4](#_heading=h.4i7ojhp)

[2. Sơ đồ xây dựng dữ liệu 4](#_heading=h.2xcytpi)

[2.1. Định nghĩa các lược đồ dữ liệu 4](#_heading=h.1ci93xb)

[2.2. Lược đồ kho dữ liệu 5](#_heading=h.3whwml4)

[2.3. Mô tả chi tiết các bảng dữ liệu 5](#_heading=h.qsh70q)

[Chương 2. QUÁ TRÌNH TÍCH HỢP DỮ LIỆU – SSIS 9](#_heading=h.ihv636)

[1. Chuẩn bị công cụ 9](#_heading=h.32hioqz)

[2. Tạo database trong Microsoft SQL Server Management Studio 9](#_heading=h.1hmsyys)

[3. Quá trình SSIS 10](#_heading=h.vx1227)

[3.1. Khởi tạo Project 10](#_heading=h.3fwokq0)

[3.2. Khởi tạo Data Flow Task 11](#_heading=h.2u6wntf)

[Chương 3. QUÁ TRÌNH SSAS 15](#_heading=h.3l18frh)

[1. Chuẩn bị công cụ 15](#_heading=h.206ipza)

[2. Quá trình SSAS 16](#_heading=h.1egqt2p)

[2.1. Cài đặt ràng buộc cho database 16](#_heading=h.3ygebqi)

[2.2. Cài đặt kết nối đến database 16](#_heading=h.sqyw64)

[2.3. Cài đặt kết nối Data Sources 17](#_heading=h.1rvwp1q)

[2.4. Cài đặt kết nối Data Source Views 18](#_heading=h.3q5sasy)

[2.5. Tạo Cube và Dimensions 21](#_heading=h.3hv69ve)

[2.6. Deploy quá trình 25](#_heading=h.39kk8xu)

[Chương 4. PHÂN TÍCH SSAS VÀ TRUY VẤN MDX 26](#_heading=h.2nusc19)

[1. Quá trình chuẩn bị 26](#_heading=h.1302m92)

[2. Thực hiện phân tích SSAS và truy vấn MDX 27](#_heading=h.haapch)

[2.1. Thống kê phân tích tổng doanh thu (Amount) theo năm 27](#_heading=h.319y80a)

[2.2. Thống kê tổng lợi nhuận (Profit) theo loại sản phẩm (Coffee Type) 28](#_heading=h.upglbi)

[2.3. Thống kê số lượng bán (Quantity) theo quốc gia (Country) 30](#_heading=h.184mhaj)

[2.4. Thống kê lợi nhuận (Profit) theo thành phố (City) trong năm 2022 31](#_heading=h.36ei31r)

[2.5. Thống kê doanh thu (Amount) theo phân loại khách hàng Loyalty 33](#_heading=h.zu0gcz)

[2.6. Thống kê doanh thu (Amount) theo từng tháng trong năm 34](#_heading=h.2y3w247)

[2.7. Thống kê 5 Quốc gia (Country) có doanh thu (Amount) cao nhất 36](#_heading=h.rjefff)

[2.8. Thống kê số lượng sản phẩm bán ra (Quantity) theo các ngày trong tuần (Day of Week) 37](#_heading=h.2pta16n)

[2.9. Thống kê sản lượng loại cà phê (Coffee Type) và cách rang (Roast Type) được khách hàng yêu thích 39](#_heading=h.j8sehv)

[2.10. Thống kê sản phẩm có doanh thu (Profit) cao nhất 40](#_heading=h.2hio093)

[2.11. Thống kê doanh thu mỗi thành phố (City) của mỗi quốc gia (Country) 42](#_heading=h.2uxtw84)

[Chương 5. QUÁ TRÌNH REPORT POWER BI (SSRS) 43](#_heading=h.2981zbj)

[1. Đổ dữ liệu từ SSAS vào Power BI 43](#_heading=h.odc9jc)

[2. Thực hiện SSRS 45](#_heading=h.3ls5o66)

[2.1. Thống kê tổng số doanh thu (Amount) theo năm (Year) 45](#_heading=h.20xfydz)

[2.2. Thống kê tổng số lợi nhuận (Profit) theo loại sản phẩm (Coffee Type) 46](#_heading=h.302dr9l)

[2.3. Thống kê số lượng bán (Quantity) theo quốc gia (Country) 46](#_heading=h.3z7bk57)

[2.4. Thống kê lợi nhuận (Profit) theo vùng (City) trong năm 2022 47](#_heading=h.thw4kt)

[2.5. Thống kê tổng số doanh thu (Amount) của khách hàng có Loyalty Card 47](#_heading=h.1smtxgf)

[2.6. Thống kê doanh thu (Amount) theo từng tháng mỗi năm 48](#_heading=h.2rrrqc1)

[2.7. Thống kê top 5 quốc gia (Country) có doanh thu (Amount) cao nhất 48](#_heading=h.3qwpj7n)

[2.8. số lượng sản phẩm bán ra (Quantity) nhiều nhất theo các ngày trong tuần (Day of Week) 49](#_heading=h.l7a3n9)

[2.9. Thống kê khách hàng thích loại cà phê (Coffee Type) và cách rang (Roast Type) nào 49](#_heading=h.1kc7wiv)

[*2.10. Thống kê sản phẩm mang lại doanh thu (Profit) nhiều nhất* 50](#_heading=h.2jh5peh)

[2.11. Thống kê doanh thu của từng quốc gia (Country) cho từng thành phố (City) 50](#_heading=h.3im3ia3)

[Chương 6. QUÁ TRÌNH KHAI PHÁ DỮ LIỆU 51](#_heading=h.4hr1b5p)

[1. Giới thiệu 51](#_heading=h.2wwbldi)

[2. Các bước thực hiện theo từng thuật toán 52](#_heading=h.1c1lvlb)

[2.1. Random Forest 52](#_heading=h.3w19e94)

[2.2. Moving Averages 53](#_heading=h.49gfa85)

[2.3. LSTM 54](#_heading=h.3nqndbk)

[2.4. Linear Regression 56](#_heading=h.3fg1ce0)

[3. Trực quan hóa 57](#_heading=h.n5rssn)

[3.1. Random Forest 57](#_heading=h.375fbgg)

[3.2. Moving Averages 58](#_heading=h.46ad4c2)

[3.3. LSTM 58](#_heading=h.10kxoro)

[3.4. Linear Regression 59](#_heading=h.1zpvhna)

[4. Kết quả thực hiện 59](#_heading=h.2yutaiw)

[Chương 7. KẾT LUẬN 61](#_heading=h.3xzr3ei)

[1. Kết quả đạt được 61](#_heading=h.2d51dmb)

[2. Hạn chế 61](#_heading=h.sabnu4)

[3. Phân công công việc 62](#_heading=h.3c9z6hx)

# DANH MỤC BẢNG

[Bảng 1.1: Thuộc tính kho dữ liệu Order 1](#_heading=h.35nkun2)

[Bảng 1.2: Thuộc tính kho dữ liệu Customer 2](#_heading=h.1ksv4uv)

[Bảng 1.3: Thuộc tính kho dữ liệu Product 2](#_heading=h.44sinio)

[Bảng 1.4: Thuộc tính kho dữ liệu Date 3](#_heading=h.2jxsxqh)

[Bảng 1.5: Mô tả chi tiết thuộc tính Coffee 3](#_heading=h.3j2qqm3)

[Bảng 1.6: Mô tả chi tiết thuộc tính Roast type 3](#_heading=h.1y810tw)

[Bảng 1.7: Mô tả chi tiết bảng dữ liệu Fact 6](#_heading=h.3as4poj)

[Bảng 1.8: Mô tả chi tiết bảng dữ liệu Dim\_Product 6](#_heading=h.1pxezwc)

[Bảng 1.9: Mô tả chi tiết bảng dữ liệu Dim\_Customer 7](#_heading=h.49x2ik5)

[Bảng 1.10: Mô tả chi tiết bảng dữ liệu Dim\_Time 7](#_heading=h.2p2csry)

[Bảng 1.11: Mô tả chi tiết bảng dữ liệu Dim\_Country 8](#_heading=h.147n2zr)

[Bảng 1.12: Mô tả chi tiết bảng dữ liệu Dim\_City 8](#_heading=h.3o7alnk)

[Bảng 1.13: Mô tả chi tiết bảng dữ liệu Dim\_CoffeeType 8](#_heading=h.23ckvvd)

[Bảng 6.1: Kết quả thông số sai lệch theo từng mô hình 59](#_heading=h.1e03kqp)

DANH MỤC HÌNH ẢNH

# 

[Hình 1.1: Lược đồ kho dữ liệu Coffee theo snowflake schema 5](#_heading=h.2bn6wsx)

[Hình 2.1: Giao diện kết nối Database Engine trong Microsoft SQL Server 9](#_heading=h.41mghml)

[Hình 2.2: Giao diện tạo database coffee trong Microsoft SQL Server 9](#_heading=h.2grqrue)

[Hình 2.3: Giao diện khởi tạo Integration Services Project trong Visual Studio 2022 10](#_heading=h.1v1yuxt)

[Hình 2.4: Giao diện đặt tên cho Project trong Visual Studio 2022 10](#_heading=h.4f1mdlm)

[Hình 2.5: Giao diện tạo kết nối với database trong Visual Studio 2022 11](#_heading=h.19c6y18)

[Hình 2.6: Giao diện tạo Data Flow Task trong Visual Studio 2022 11](#_heading=h.3tbugp1)

[Hình 2.7: Giao diện import dữ liệu cho quá trình SSIS trong Visual Studio 2022 12](#_heading=h.28h4qwu)

[Hình 2.8: Giao diện sơ lược quá trình SSIS trong Visual Studio 2022 12](#_heading=h.37m2jsg)

[Hình 2.9: Giao diện cấu hình cho các Node Sort trong Visual Studio 2022 13](#_heading=h.1mrcu09)

[Hình 2.10: Giao diện tạo OLE DB Destination trong Visual Studio 2022 13](#_heading=h.46r0co2)

[Hình 2.11: Giao diện xử lý câu lệnh SQL Execute Task trong Visual Studio 2022 14](#_heading=h.2lwamvv)

[Hình 2.12: Giao diện kết quả thực hiện quá trình SSIS 14](#_heading=h.111kx3o)

[Hình 3.1: Giao diện tạo Analysis Services Multidimensional Project trong Visual Studio 2022 15](#_heading=h.4k668n3)

[Hình 3.2: Giao diện đặt tên cho SSAS Project trong Visual Studio 2022 15](#_heading=h.2zbgiuw)

[Hình 3.3: Giao diện thực thi ràng buộc cho database 16](#_heading=h.2dlolyb)

[Hình 3.4: Giao diện tạo kết nối database cho quá trình SSAS trong Visual Studio 2022 16](#_heading=h.3cqmetx)

[Hình 3.5: Giao diện tạo kết nối Data Sources trong Visual Studio 2022 17](#_heading=h.4bvk7pj)

[Hình 3.6: Giao diện đăng nhập bằng tài khoản Window trong Visual Studio 2022 17](#_heading=h.2r0uhxc)

[Hình 3.7: Giao diện tạo và đặt tên cho Data source 18](#_heading=h.1664s55)

[Hình 3.8: Giao diện chọn cài đặt Data Source trong Visual Studio 2022 18](#_heading=h.25b2l0r)

[Hình 3.9: Giao diện tạo các logical relationships trong Visual Studio 2022 19](#_heading=h.kgcv8k)

[Hình 3.10: Giao diện kéo các bảng liên quan đến bảng Fact 19](#_heading=h.1jlao46)

[Hình 3.11: Giao diện chọn các bảng liên quan đến bảng Dim\_customer 20](#_heading=h.43ky6rz)

[Hình 3.12: Giao diện chọn các bảng liên quan đến bảng Dim\_product 20](#_heading=h.2iq8gzs)

[Hình 3.13: Giao diện đặt tên cho Data Source View 21](#_heading=h.xvir7l)

[Hình 3.14: Giao diện khởi tạo Cube 21](#_heading=h.1x0gk37)

[Hình 3.15: Giao diện chọn Measure Group Table 22](#_heading=h.4h042r0)

[Hình 3.16: Giao diện chọn các Dimensions 22](#_heading=h.2w5ecyt)

[Hình 3.17: Kiểm tra lại các thông tin trước khi tạo Cube 23](#_heading=h.1baon6m)

[Hình 3.18: Giao diện kết quả cube đã tạo 23](#_heading=h.3vac5uf)

[Hình 3.19: Giao diện kéo thêm các thuộc tính chưa xuất hiện sang Attributes 24](#_heading=h.2afmg28)

[Hình 3.20: Giao diện tạo thêm các New Dimension 24](#_heading=h.pkwqa1)

[Hình 3.21: Giao diện chọn Process thực thi quá trình deployment 25](#_heading=h.1opuj5n)

[Hình 3.22: Giao diện deployment thành công 25](#_heading=h.48pi1tg)

[Hình 4.1: Giao diện chuẩn bị thực hiện Pivot trong Microsoft Excel 26](#_heading=h.3mzq4wv)

[Hình 4.2: Giao diện kết nối Analysis Service chuẩn bị truy vấn MDX 26](#_heading=h.2250f4o)

[Hình 4.3: Kết quả truy vấn MDX thống kê phân tích tổng doanh thu theo năm 27](#_heading=h.1gf8i83)

[Hình 4.4: Kết quả Pivot table thống kê phân tích tổng doanh thu theo năm 27](#_heading=h.40ew0vw)

[Hình 4.5: Kết quả Manual SSAS thống kê phân tích tổng doanh thu theo năm 28](#_heading=h.2fk6b3p)

[Hình 4.6: Kết quả truy vấn MDX thống kê tổng lợi nhuận theo loại sản phẩm 28](#_heading=h.3ep43zb)

[Hình 4.7: Kết quả Pivot Table thống kê tổng lợi nhuận theo loại sản phẩm 29](#_heading=h.4du1wux)

[Hình 4.8: Kết quả Manual SSAS thống kê tổng lợi nhuận theo loại sản phẩm 29](#_heading=h.2szc72q)

[Hình 4.9: Kết quả truy vấn MDX thống kê số lượng bán theo quốc gia 30](#_heading=h.3s49zyc)

[Hình 4.10: Kết quả Pivot table thống kê số lượng bán theo quốc gia 30](#_heading=h.279ka65)

[Hình 4.11: Kết quả Manual SSAS thống kê số lượng bán theo quốc gia 31](#_heading=h.meukdy)

[Hình 4.12: Kết quả truy vấn MDX thống kê lợi nhuận theo thành phố năm 2022 31](#_heading=h.1ljsd9k)

[Hình 4.13: Kết quả Pivot table thống kê lợi nhuận theo thành phố năm 2022 32](#_heading=h.45jfvxd)

[Hình 4.14: Kết quả Manual SSAS thống kê lợi nhuận theo thành phố năm 2022 32](#_heading=h.2koq656)

[Hình 4.15: Kết quả truy vấn MDX thống kê doanh thu theo phân loại khách hàng 33](#_heading=h.3jtnz0s)

[Hình 4.16: Kết quả Pivot table thống kê doanh thu theo phân loại khách hàng 33](#_heading=h.1yyy98l)

[Hình 4.17: Kết quả Manual SSAS thống kê doanh thu theo phân loại khách hàng 34](#_heading=h.4iylrwe)

[Hình 4.18: Kết quả truy vấn MDX thống kê doanh thu theo từng tháng trong năm 34](#_heading=h.1d96cc0)

[Hình 4.19: Kết quả Pivot table thống kê doanh thu theo từng tháng trong năm 35](#_heading=h.3x8tuzt)

[Hình 4.20: Kết quả Manual SSAS thống kê doanh thu theo từng tháng trong năm 35](#_heading=h.2ce457m)

[Hình 4.21: Kết quả truy vấn MDX thống kê 5 Quốc gia có doanh thu cao nhất 36](#_heading=h.3bj1y38)

[Hình 4.22: Kết quả Pivot table thống kê 5 Quốc gia có doanh thu cao nhất 36](#_heading=h.1qoc8b1)

[Hình 4.23: Kết quả Manual SSAS thống kê 5 Quốc gia có doanh thu cao nhất 37](#_heading=h.4anzqyu)

[Hình 4.24: Kết quả MDX thống kê số lượng sản phẩm bán ra theo ngày trong tuần 37](#_heading=h.14ykbeg)

[Hình 4.25: Kết quả Pivot table thống kê số lượng sản phẩm bán ra theo các ngày trong tuần 38](#_heading=h.3oy7u29)

[Hình 4.26: Kết quả Manual SSAS thống kê số lượng sản phẩm bán ra theo các ngày trong tuần 38](#_heading=h.243i4a2)

[Hình 4.27: Kết quả MDX thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích 39](#_heading=h.338fx5o)

[Hình 4.28: Kết quả Pivot table thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích 39](#_heading=h.1idq7dh)

[Hình 4.29: Kết quả Manual SSAS thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích 40](#_heading=h.42ddq1a)

[Hình 4.30: Kết quả MDX thống kê sản phẩm có doanh thu cao nhất 40](#_heading=h.3gnlt4p)

[Hình 4.31: Kết quả Pivot table thống kê sản phẩm có doanh thu cao nhất 41](#_heading=h.1vsw3ci)

[Hình 4.32: Kết quả Manual SSAS thống kê sản phẩm có doanh thu cao nhất 41](#_heading=h.4fsjm0b)

[Hình 4.33: Kết quả MDX thống kê doanh thu mỗi thành phố của quốc gia 42](#_heading=h.1a346fx)

[Hình 4.34: Kết quả Manual SSASthống kê doanh thu mỗi thành phố của quốc gia 42](#_heading=h.3u2rp3q)

[Hình 5.1: Giao diện phần mềm Power BI 43](#_heading=h.38czs75)

[Hình 5.2: Giao diện chọn công cụ Analysis Services trong Power BI 43](#_heading=h.1nia2ey)

[Hình 5.3: Giao diện chọn Server từ Microsoft SQL Server 44](#_heading=h.47hxl2r)

[Hình 5.4: Giao diện chọn Cube để lấy dữ liệu từ quá trình SSAS 44](#_heading=h.2mn7vak)

[Hình 5.5: Giao diện đổ dữ liệu từ Analysis Service hoàn tất 45](#_heading=h.11si5id)

[Hình 5.6: Giao diện thống kê phân tích tổng doanh thu theo năm 45](#_heading=h.4kx3h1s)

[Hình 5.7: Giao diện thống kê tổng lợi nhuận theo loại sản phẩm 46](#_heading=h.1f7o1he)

[Hình 5.8: Giao diện thống kê số lượng bán theo quốc gia 46](#_heading=h.2eclud0)

[Hình 5.9: Giao diện thống kê lợi nhuận theo thành phố năm 2022 47](#_heading=h.3dhjn8m)

[Hình 5.10: Giao diện thống kê doanh thu theo phân loại khách hàng 47](#_heading=h.4cmhg48)

[Hình 5.11: Giao diện thống kê doanh thu theo từng tháng trong năm 48](#_heading=h.16x20ju)

[Hình 5.12: Giao diện thống kê 5 Quốc gia có doanh thu cao nhất 48](#_heading=h.261ztfg)

[Hình 5.13: Giao diện thống kê 49](#_heading=h.356xmb2)

[Hình 5.14: Giao diện thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích 49](#_heading=h.44bvf6o)

[Hình 5.15: Giao diện thống kê sản phẩm có doanh thu cao nhất 50](#_heading=h.ymfzma)

[Hình 5.16: Giao diện thống kê doanh thu mỗi thành phố của quốc gia 50](#_heading=h.1xrdshw)

[Hình 6.1: Import và sắp xếp dữ liệu 52](#_heading=h.2b6jogx)

[Hình 6.2: Chia tập train, test theo tỉ lệ 7:3 52](#_heading=h.qbtyoq)

[Hình 6.3: Truyền các tham số đầu vào cho mô hình Random Forest 52](#_heading=h.3abhhcj)

[Hình 6.4: Dự đoán doanh thu 1 năm sau từ thời điểm gần nhất 53](#_heading=h.1pgrrkc)

[Hình 6.6: Chia tập train, test theo tỉ lệ 7:3 53](#_heading=h.2olpkfy)

[Hình 6.8: Dự đoán doanh thu 1 năm sau từ thời điểm gần nhất 54](#_heading=h.13qzunr)

[Hình 6.9: Import và nhóm cột Amount theo ngày 54](#_heading=h.22vxnjd)

[Hình 6.10: Scale tỉ lệ và chia tập train, test theo tỉ lệ 7:3 54](#_heading=h.i17xr6)

[Hình 6.11: chuẩn bị dữ liệu cho mô hình LSTM 55](#_heading=h.320vgez)

[Hình 6.12: tạo dữ liệu đầu vào và đầu ra cho mô hình LSTM 55](#_heading=h.1h65qms)

[Hình 6.13: Tạo mô hình tuần tự và huấn luyện mô hình 55](#_heading=h.415t9al)

[Hình 6.14: Chuyển đổi ngược kết quả dự đoán về thang đo ban đầu 55](#_heading=h.2gb3jie)

[Hình 6.15: Dự đoán doanh thu 1 năm sau từ thời điểm gần nhất 56](#_heading=h.vgdtq7)

[Hình 6.16: Import và nhóm cột Amount theo ngày 56](#_heading=h.1ulbmlt)

[Hình 6.17: Tạo cột Days để thực hiện mô hình hồi quy tuyến tính 56](#_heading=h.4ekz59m)

[Hình 6.18: Chia tập train, test theo tỉ lệ 7:3 56](#_heading=h.2tq9fhf)

[Hình 6.19: Tạo các đặc trưng đa thức bậc 2 57](#_heading=h.18vjpp8)

[Hình 6.20: Tạo và huấn luyện mô hình Linear Regression 57](#_heading=h.3sv78d1)

[Hình 6.21: Chuyển đổi dữ liệu tương lại thành dạng đa thức 57](#_heading=h.280hiku)

[Hình 6.22: Kết quả dự đoán mô hình Random Forest 57](#_heading=h.1maplo9)

[Hình 6.23: Kết quả dự đoán mô hình Moving Average 58](#_heading=h.2lfnejv)

[Hình 6.24: Kết quả dự đoán mô hình LSTM 58](#_heading=h.3kkl7fh)

[Hình 6.25: Kết quả dự đoán mô hình Linear Regression 59](#_heading=h.4jpj0b3)

# Chương 1. GIỚI THIỆU TỔNG QUAN

## 1. Mô tả bộ dữ liệu

### 1.1. Mô tả bộ dữ liệu

Bộ dữ liệu **Coffee Bean Sales Raw Dataset** là bộ dữ liệu doanh số bán cà phê của một nơi tại Mỹ được công bố trên trang Kaggle.com cập nhập vào 1 năm trước

Thông qua bộ dữ liệu, người dùng có thể biết được thông tin chi tiết đơn hàng, thông tin của khách hàng mua hàng, thông tin chi tiết bao gồm thành phố và quốc gia của khách hàng, thông tin, giá sản phẩm, số lượng mua hàng, lợi nhuận và doanh thu…

Dataset gốc có 3 sheet, sheet ‘**orders’** có 5 cột và 1000 dòng, sheet ‘**customers’** có 9 cột và 1000 dòng và sheet ‘**products’** có 7 cột và 48 dòng

Nguồn dữ liệu: [Coffee Bean Sales Raw Dataset](https://www.kaggle.com/datasets/saadharoon27/coffee-bean-sales-raw-dataset)

### 1.2. Thuộc tính kho dữ liệu

* **Order**

**Bảng 1.1: Thuộc tính kho dữ liệu Order**

| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| --- | --- | --- | --- |
| 1 | Order ID | nvarchar(150) | Mã số định danh cho đơn đặt hàng |
| 2 | Order Date | datetime | Ngày đơn đặt hàng |
| 3 | Customer ID | nvarchar(255) | Mã số định danh cho khách hàng |
| 4 | Product ID | nvarchar(75) | Mã số định danh cho sản phẩm cà phê |
| 5 | Quantity | float | Số lương bán ra theo sản phẩm của đơn hàng |
| 6 | Amount | float | Doanh thu theo sản phẩm bán ra của đơn hàng |
| 7 | Profit | float | Lợi nhuận theo sản phẩm bán ra của đơn hàng |

* **Customer**

**Bảng 1.2: Thuộc tính kho dữ liệu Customer**

| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| --- | --- | --- | --- |
| 1 | Customer ID | nvarchar(255) | Mã số định danh cho khách hàng |
| 2 | Customer Name | nvarchar(255) | Tên khách hàng |
| 3 | Email | nvarchar(255) | Email của khách hàng |
| 4 | Phone Number | nvarchar(255) | Số điện thoại |
| 5 | Adddresss Line 1 | nvarchar(255) | Địa chỉ |
| 6 | City | nvarchar(255) | Thành phố sinh sống |
| 7 | Country | nvarchar(255) | Nước sinh sống |
| 8 | Postcode | float | Mã bưu chính |
| 9 | Loyalty Card | nvarchar(255) | Khách hàng có thẻ thành viên không |

* **Product**

**Bảng 1.3: Thuộc tính kho dữ liệu Product**

| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| --- | --- | --- | --- |
| 1 | Product ID | nvarchar(75) | Mã số định danh cho sản phẩm cà phê |
| 2 | Coffee Type | nvarchar(255) | Loại cà phê |
| 3 | Roast Type | nvarchar(255) | Kiểu rang của cà phê |
| 4 | Size | float | Kích cỡ sản phẩm (theo kg) |
| 5 | Unit Price | float | Giá theo đơn vị sản phẩm |
| 6 | Price per 100g | float | Giá theo 100g của sản phẩm |
| 7 | Profit | float | Lợi nhuận của đơn vị sản phẩm |

* **Date**

**Bảng 1.4: Thuộc tính kho dữ liệu Date**

| **STT** | **Thuộc tính** | **Kiểu dữ liệu** | **Ý nghĩa** |
| --- | --- | --- | --- |
| 1 | Order Date | datetime | Ngày đơn đặt hàng |
| 2 | Day | float | Ngày |
| 3 | Month | float | Tháng |
| 4 | Year | float | Năm |
| 5 | Day of Week | float | Thứ mấy dựa trên Order Date |
| 6 | Quarter | float | Quý |

### 1.3. Mô tả chi tiết thuộc tính

Thuộc tính **Coffee:** phân loại cà phê theo chủng loại cà phê

**Bảng 1.5: Mô tả chi tiết thuộc tính Coffee**

| **Coffee Type** | | |
| --- | --- | --- |
| **STT** | **Giá trị** | **Ý nghĩa** |
| 1 | Ara | Arabica |
| 2 | Exc | Excelsa |
| 3 | Lib | Liberica |
| 4 | Rob | Robusta |

Thuộc tính **Roast Type:** phân loại cà phê theo cách rang cà phê

**Bảng 1.6: Mô tả chi tiết thuộc tính Roast type**

| **Roast Type** | | |
| --- | --- | --- |
| **STT** | **Giá trị** | **Ý nghĩa** |
| 1 | D | Dark (Rang đậm) |
| 2 | L | Light (Rang nhẹ) |
| 3 | M | M (Rang vừa) |

### 1.4. Tiền xử lý bộ dữ liệu

Nhóm thực hiện tiền xử lý bằng cách thêm các cột measure như ‘Amount’, ‘Profit’, loại 1 số dòng không cần thiết

Dataset sau khi được xử lý sẽ có 4 sheet, sheet ‘orders’ có 7 cột và 999 dòng, sheet ‘customers’ có 9 cột và 1000 dòng, sheet ‘products’ có 7 cột và 48 dòng và sheet ‘date’ có 6 cột và 689 dòng

## 2. Sơ đồ xây dựng dữ liệu

### 2.1. Định nghĩa các lược đồ dữ liệu

Các lược đồ dữ liệu đa chiều:

- Lược đồ đa chiều được thiết kế đặc biệt để mô hình hoá các hệ thống kho dữ liệu. Các lược đồ được thiết kế để giải quyết các nhu cầu của cơ sở dữ liệu rất lớn được thiết kế cho mục đích phân tích dữ liệu (OLAP). Có 2 loại lược đồ đa chiều:

#### *2.1.1.* Lược đồ hình sao (Star Schema)

Lược đồ hình sao là loại được đồ kho dữ liệu đơn giản nhất. Nó được gọi là được đồ hình sao vì cấu trúc giống như một ngôi sao. Trong lược đồ hình sao, tâm của ngôi sao có thể có một bảng sự kiện (Fact) và có các bảng chiều (Dimensions) được liên kết. Nó còn được gọi là Star Join Schema và được tối ưu hoá để truy vấn các tập dữ liệu lớn.

* Ưu điểm:
* Các truy vấn đơn giản hơn – login nối lược đồ sao thường đơn giản hơn logic nối được yêu cầu truy xuất dữ liệu từ lược đồ giao dịch được chuẩn hoá cao.
* Tăng hiệu suất truy vấn – các lược đồ sao có thể cung cấp các cải tiến hiệu suất cho các ứng dụng báo cáo chỉ đọc khi so sánh với các lượt đồ được chuẩn hoá cao.
* Tổng hợp nhanh – các truy vấn đơn giản hơn đối với lược đồ sao có thể dẫn dến hiệu suất được cải thiện cho các hoạt động tổng hợp.
* Nhược điểm:
* Tính toàn vẹn dữ liệu không được thực thi tốt vì nó không ở trạng thái chuẩn hóa cao.
* Lược đồ hình sao không linh hoạt về mặt nhu cầu phân tích như mô hình dữ liệu chuẩn hóa.

#### 2.1.2. Lược đồ bông tuyết (Snowflake Schema)

Lược đồ Bông tuyết là một phần mở rộng của Lược đồ hình sao và nó thêm các kích thước bổ sung. Nó được gọi là bông tuyết vì sơ đồ của nó giống như một Bông tuyết.

Ưu điểm:

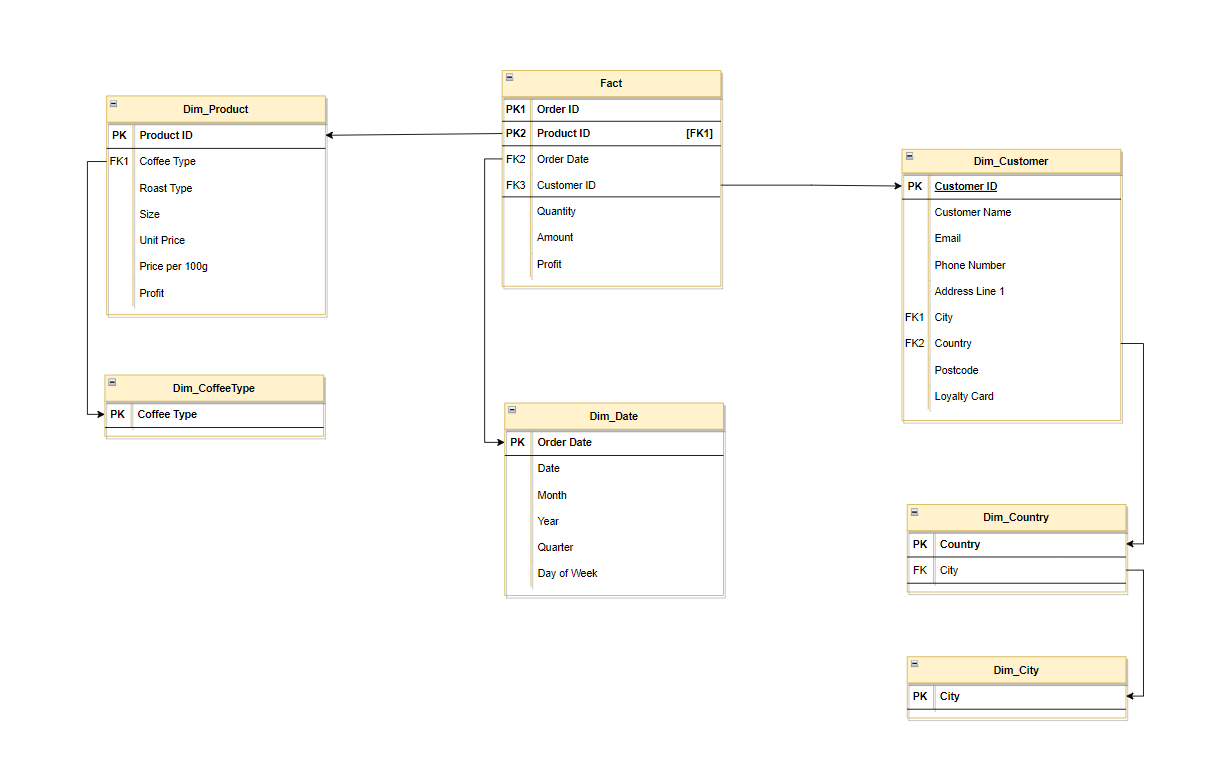
Lược đồ bông tuyết nằm trong cùng một họ với mô hình logic hình sao. Lược đồ hình sao được coi là một trường hợp đặc biệt của lược đồ bông tuyết.

* Ưu điểm:
* Một số công cụ mô hình hóa cơ sở dữ liệu đa chiều (OLAP) được tối ưu hóa cho các lược đồ bông tuyết.
* Đơn giản hóa các thuộc tính dẫn đến sự tiết kiệm, nhưng đánh đổi là sự phức tạp bổ sung trong các truy vấn nguồn.
* Một số chiều được phân cấp để thể hiện rõ ràng dạng chuẩn của bảng chiều.
* Nhược điểm:

Mức chuẩn hóa thuộc tính bổ sung thêm độ phức tạp cho các phép truy vấn nguồn so với lược đồ hình sao.

### 2.2. Lược đồ kho dữ liệu

* Nhóm xây dựng kho dữ liệu theo snowflake schema



**Hình 1.1: Lược đồ kho dữ liệu Coffee theo snowflake schema**

### 2.3. Mô tả chi tiết các bảng dữ liệu

* **Bảng Fact**

**Bảng 1.7: Mô tả chi tiết bảng dữ liệu Fact**

| **Fact** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| Order ID | nvarchar(150) | Khóa chính |
| Order Date | datetime | Khóa ngoại |
| Customer ID | nvarchar(255) | Khóa ngoại |
| Product ID | nvarchar(75) | Khóa chính, khóa ngoại |
| Quantity | float |  |
| Amount | float |  |
| Profit | float |  |

* **Bảng Dim\_Product**

**Bảng 1.8: Mô tả chi tiết bảng dữ liệu Dim\_Product**

| **Dim\_Product** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| Product ID | nvarchar(75) | Khóa chính |
| Coffee Type | nvarchar(255) | Khóa ngoại |
| Roast Type | nvarchar(255) |  |
| Size | float |  |
| Unit Price | float |  |
| Price per 100g | float |  |
| Profit | float |  |

* **Bảng Dim\_Customer**

**Bảng 1.9: Mô tả chi tiết bảng dữ liệu Dim\_Customer**

| **Dim\_Customer** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| Cusotmer ID | nvarchar(255) | Khóa chính |
| Customer Name | nvarchar(255) |  |
| Email | nvarchar(255) |  |
| Phone Number | nvarchar(255) |  |
| Address Line 1 | nvarchar(255) |  |
| City | nvarchar(255) | Khóa ngoại |
| Country | nvarchar(255) | Khóa ngoại |
| Postcode | float |  |
| Loyalty Card | nvarchar(255) |  |

* **Bảng Dim\_Time**

**Bảng 1.10: Mô tả chi tiết bảng dữ liệu Dim\_Time**

| **Dim\_Time** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| Order Date | nvarchar(255) | Khóa chính |
| Day | datetime |  |
| Month | nvarchar(255) |  |
| Year | nvarchar(255) |  |
| Quarter | float |  |
| Day of Week | float |  |

* **Bảng Dim\_Country**

**Bảng 1.11: Mô tả chi tiết bảng dữ liệu Dim\_Country**

| **Dim\_Country** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| Country | nvarchar(255) | Khóa chính |

* **Bảng Dim\_City**

**Bảng 1.12: Mô tả chi tiết bảng dữ liệu Dim\_City**

| **Dim\_City** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| City | nvarchar(255) | Khóa chính |

* **Bảng Dim\_CoffeeType**

**Bảng 1.13: Mô tả chi tiết bảng dữ liệu Dim\_CoffeeType**

| **Dim\_CoffeeType** | | |
| --- | --- | --- |
| **Tên thuộc tính** | **Kiểu dữ liệu** | **Ràng buộc** |
| Coffee Type | nvarchar(255) | Khóa chính |

# Chương 2. QUÁ TRÌNH TÍCH HỢP DỮ LIỆU – SSIS

## 1. Chuẩn bị công cụ

* Tải Microsoft Visual Studio 2022.
* Tải Microsoft SQL Server Management Studio.
* Tải công cụ SQL Server Data Tools cho phiên bản Visual Studio 2022.

## 2. Tạo database trong Microsoft SQL Server Management Studio

* Bước 1: Kết nối Database Engine

A screenshot of a computer

Description automatically generated

**Hình 2.1: Giao diện kết nối Database Engine trong Microsoft SQL Server**

* Bước 2: Dùng lệnh tạo database trong cửa sổ Query

A screenshot of a computer

Description automatically generated

**Hình 2.2: Giao diện tạo database coffee trong Microsoft SQL Server**

## 3. Quá trình SSIS

### 3.1. Khởi tạo Project

* Khởi tạo **Integration Services Project** sau đó nhấn **Next**

A screenshot of a computer

Description automatically generated

**Hình 2.3: Giao diện khởi tạo Integration Services Project trong Visual Studio 2022**

* Đặt tên cho Project, nơi lưu trữ file và nhấn Create

A screenshot of a computer

Description automatically generated

**Hình 2.4: Giao diện đặt tên cho Project trong Visual Studio 2022**

### 3.2. Khởi tạo Data Flow Task

#### 3.2.1. Tạo kết nối với database

A computer screen with a white box

Description automatically generated

**Hình 2.5: Giao diện tạo kết nối với database trong Visual Studio 2022**

#### 3.2.2. Khởi tạo Data Flow Task

Tạo một **Data Flow Task** sau đó nhấn đúp vào để chuyển sang tab **Data Flow** để thực hiện cấu hình.

A computer screen with a message box

Description automatically generated

**Hình 2.6: Giao diện tạo Data Flow Task trong Visual Studio 2022**

#### 3.2.3. Import data cho quá trình SSIS

* Bước 1: Tại tab Data Flow thực hiện tìm và kéo ra Excel Source từ SSIS Toolbox
* Nhấn đúp vào Excel Source để mở Excel Source Editortiến hành tạo mới một kết nối.

A screenshot of a computer

Description automatically generated

**Hình 2.7: Giao diện import dữ liệu cho quá trình SSIS trong Visual Studio 2022**

#### 3.2.4. Tạo bảng Fact và các bảng Dim

* Tạo Node Muticast và các Node Sort để sắp xếp dữ liệu cho các bảng Dim và thực hiện liên kết chúng với nhau

A computer screen shot of a computer

Description automatically generated

**Hình 2.8: Giao diện sơ lược quá trình SSIS trong Visual Studio 2022**

Cấu hình cho các Node Sort bằng cách nhấn đúp và chúng. Tuỳ theo Sort cho bảng Dim nào mà sẽ tích vào các thuộc tính để đưa thuộc tính đó vào các bảng tương ứng và tích vào Remove rows with duplicate sort values.

A computer screen shot of a computer

Description automatically generated

**Hình 2.9: Giao diện cấu hình cho các Node Sort trong Visual Studio 2022**

* Tại giao diện cửa sổ OLE DB Destination Editor nhấn New để tạo mới Table. Sau đó điền Script để tạo table vào và nhấn OK.

**A screenshot of a computer

Description automatically generated**

**Hình 2.10: Giao diện tạo OLE DB Destination trong Visual Studio 2022**

#### 3.2.5. Import dữ liệu vào database

* Ở SSIS Toolbox kéo thả Execute SQL Task vào Control Flow để thực thi import dữ liệu vào database
* Chọn các thông tin Connection và điền SQL Statement vào ô để xoá hết dữ liệu ở các bảng (nếu có) trước khi import dữ liệu để tránh trường hợp ghi chồng dữ liệu.

**A screenshot of a computer

Description automatically generated**

**Hình 2.11: Giao diện xử lý câu lệnh SQL Execute Task trong Visual Studio 2022**

#### 3.2.6. Kết quả quá trình SSIS

A screenshot of a computer

Description automatically generated

**Hình 2.12: Giao diện kết quả thực hiện quá trình SSIS**

# Chương 3. QUÁ TRÌNH SSAS

## 1. Chuẩn bị công cụ

* Khởi động Visual Studio 2022.
* Chọn công cụ Analysis Services Multidimensional Project, sau đó nhấn next.

A screenshot of a computer

Description automatically generated

**Hình 3.1: Giao diện tạo Analysis Services Multidimensional Project trong Visual Studio 2022**

* Đặt tên cho Project, chọn nơi lưu trữ file, và nhấn Next.

A screenshot of a computer

Description automatically generated

**Hình 3.2: Giao diện đặt tên cho SSAS Project trong Visual Studio 2022**

## 2. Quá trình SSAS

### 2.1. Cài đặt ràng buộc cho database

A screenshot of a computer

Description automatically generated

**Hình 3.3: Giao diện thực thi ràng buộc cho database**

### 2.2. Cài đặt kết nối đến database

A computer screen shot of a computer

Description automatically generated

**Hình 3.4: Giao diện tạo kết nối database cho quá trình SSAS trong Visual Studio 2022**

### 2.3. Cài đặt kết nối Data Sources

* Chọn **Create a data source based on an existing or new connection,** sau đó chọn data connections với data vừa tạo từ SSIS.

A computer screen with a message

Description automatically generated

**Hình 3.5: Giao diện tạo kết nối Data Sources trong Visual Studio 2022**

* Chọn **Use a specific Windows user name and password**, sau đó điền **Username** và **Password** là tài khoản đăng nhập Window của laptop

A computer screen with a message box

Description automatically generated

**Hình 3.6: Giao diện đăng nhập bằng tài khoản Window trong Visual Studio 2022**

* Đặt tên cho **Data** **source** là đã tạo thành công Data Source.

A computer screen with a message

Description automatically generated

**Hình 3.7: Giao diện tạo và đặt tên cho Data source**

### 2.4. Cài đặt kết nối Data Source Views

* Chọn Data Source vừa được tạo ở phần trên, sau đó nhấn Next

A computer screen with a white box

Description automatically generated

**Hình 3.8: Giao diện chọn cài đặt Data Source trong Visual Studio 2022**

* Chọn **Create logical relationships by matching columns** để tạo liên kết cho các bảng tự động với **Same name as primary key**.

A computer screen with a blue background

Description automatically generated

**Hình 3.9: Giao diện tạo các logical relationships trong Visual Studio 2022**

* Chọn bảng Fact sau đó nhấn nút > để đẩy bảng qua. Sau đó nhấn **Add Related Tables** để đẩy các bảng liên quan đến Fact qua.

A computer screen with a blue background

Description automatically generated

**Hình 3.10: Giao diện kéo các bảng liên quan đến bảng Fact**

* Chọn bảng Dim\_customer sau đó chọn Add Related Tables để đẩy bảng Dim\_city và Dim\_country qua.

A computer screen with a blue background

Description automatically generated

**Hình 3.11: Giao diện chọn các bảng liên quan đến bảng Dim\_customer**

* Chọn bảng Dim\_product sau đó chọn Add Related Tables để đẩy bảng Dim\_coffeetype qua.

A computer screen with a blue background

Description automatically generated

**Hình 3.12: Giao diện chọn các bảng liên quan đến bảng Dim\_product**

* Đặt tên cho Data Source View là đã hoàn thành quá trình tạo kết nối Data Source View.

A computer screen shot of a computer screen

Description automatically generated

**Hình 3.13: Giao diện đặt tên cho Data Source View**

### 2.5. Tạo Cube và Dimensions

* Chọn Cubes sau đó chọn New Cubes. Sau đó chọn Use an existing table để tạo cube.

A computer screen with a message box

Description automatically generated

**Hình 3.14: Giao diện khởi tạo Cube**

* Chọn **Measure Group Table** là bảng **Fact.**

A computer screen with a white screen

Description automatically generated

**Hình 3.15: Giao diện chọn Measure Group Table**

* Chọn những Dimension cần thiết → Nhấn Next.

A computer screen with a white box

Description automatically generated

**Hình 3.16: Giao diện chọn các Dimensions**

* Kiểm tra lại thông tin và nhấn Finish để hoàn tất tạo cube.

A computer screen shot of a computer

Description automatically generated

**Hình 3.17: Kiểm tra lại các thông tin trước khi tạo Cube**

* Kết quả cube đã tạo thành công.

A screenshot of a computer

Description automatically generated

**Hình 3.18: Giao diện kết quả cube đã tạo**

* Kéo thả các thuộc tính cần thiết từ Data Source View chưa xuất hiện sang Attributes

A screenshot of a computer

Description automatically generated

**Hình 3.19: Giao diện kéo thêm các thuộc tính chưa xuất hiện sang Attributes**

* Tạo thêm các Dimension chưa xuất hiện trong New Dimensions.

**A computer screen with a white box

Description automatically generated**

**Hình 3.20: Giao diện tạo thêm các New Dimension**

### 2.6. Deploy quá trình

* Chọn Process để quá trình được thực thi

A screenshot of a computer

Description automatically generated

**Hình 3.21: Giao diện chọn Process thực thi quá trình deployment**

* Sau khi xuất hiện câu lệnh Deployment Completed Successully có nghĩa là quá trình SSAS đã được deploy thành công

A screenshot of a computer

Description automatically generated

**Hình 3.22: Giao diện deployment thành công**

# Chương 4. PHÂN TÍCH SSAS VÀ TRUY VẤN MDX

## 1. Quá trình chuẩn bị

* Để thực hiện Pivot trong MS Excel, chúng ta sẽ nhấn tìm và nhấn vào nút Analyze in Excel trong Brower của Cube.

A screenshot of a computer

Description automatically generated

**Hình 4.1: Giao diện chuẩn bị thực hiện Pivot trong Microsoft Excel**

* Để thực thi các câu MDX, chúng ta vào SQL Server, sau đó kết nối với Analysis Service và tạo mới 1 cửa sổ soạn thảo MDX và thực hiện

A computer screen shot of a computer

Description automatically generated

**Hình 4.2: Giao diện kết nối Analysis Service chuẩn bị truy vấn MDX**

## 2. Thực hiện phân tích SSAS và truy vấn MDX

### 2.1. Thống kê phân tích tổng doanh thu (Amount) theo năm - Roll up

* **MDX**

A screenshot of a computer

Description automatically generated

**Hình 4.3: Kết quả truy vấn MDX thống kê phân tích tổng doanh thu theo năm -**

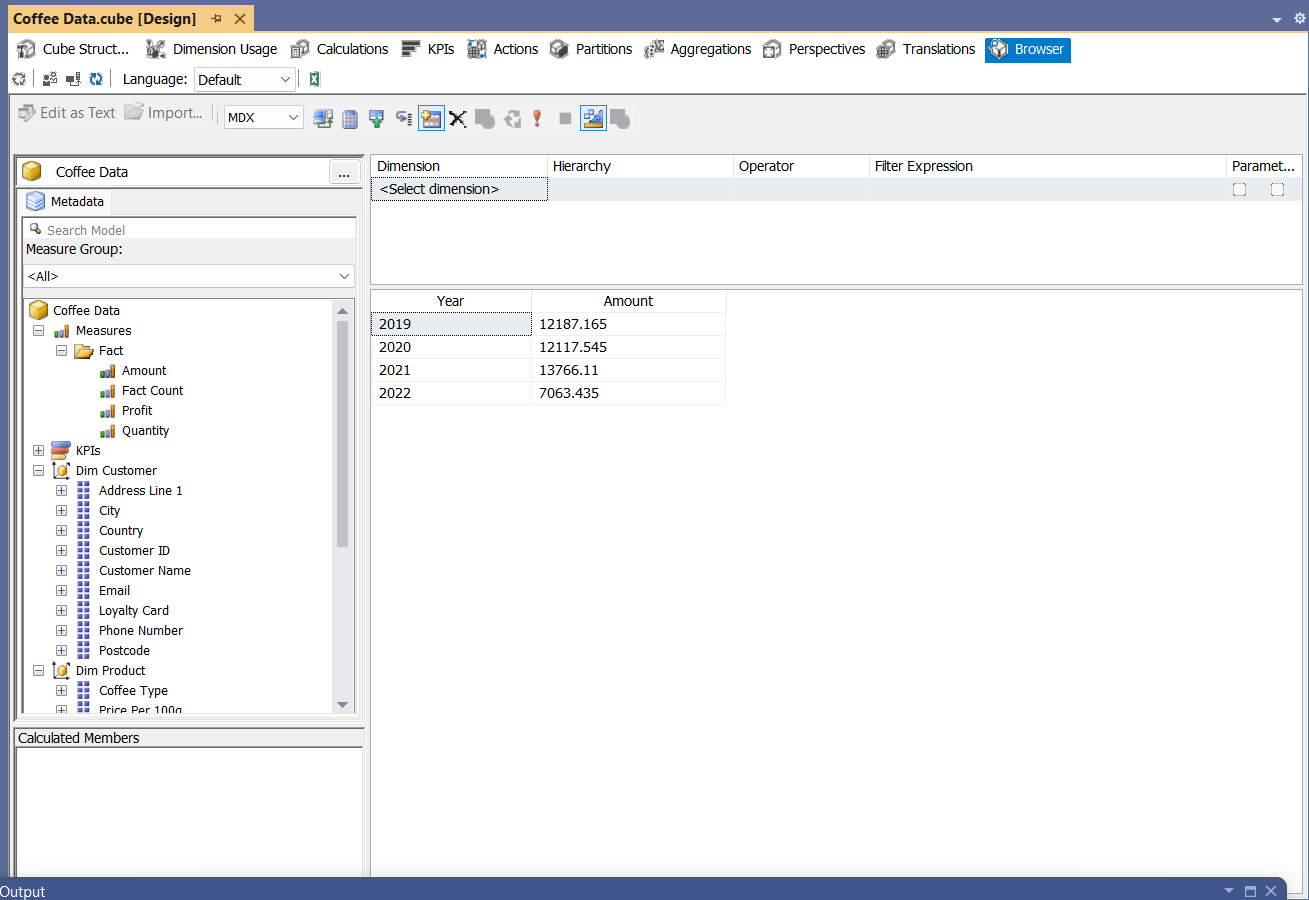
* **Pivot Table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.4: Kết quả Pivot table thống kê phân tích tổng doanh thu theo năm**

* **Manual SSAS**



**Hình 4.5: Kết quả Manual SSAS thống kê phân tích tổng doanh thu theo năm**

### 2.2. Thống kê tổng lợi nhuận (Profit) theo loại sản phẩm (Coffee Type) - Roll up

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.6: Kết quả truy vấn MDX thống kê tổng lợi nhuận theo loại sản phẩm**

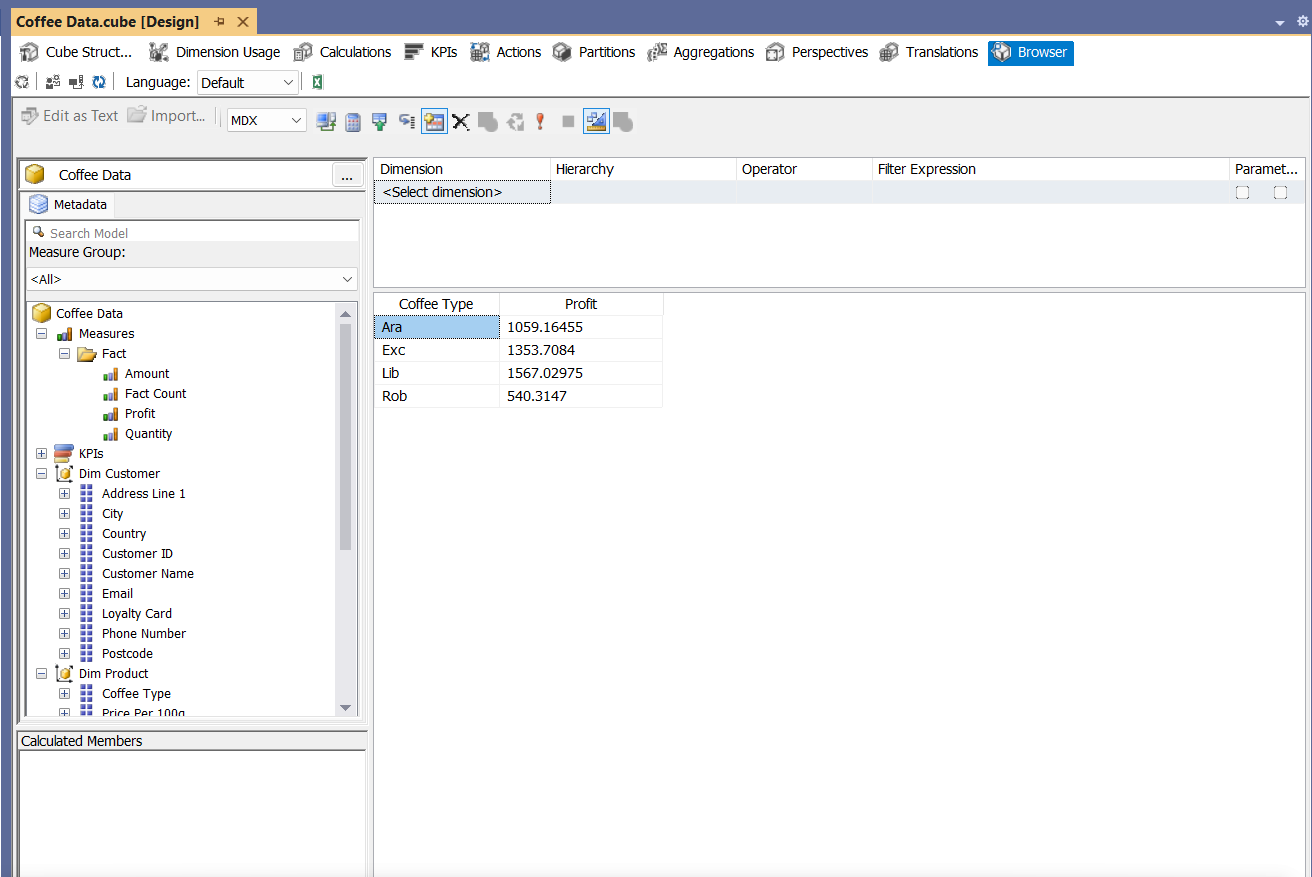
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.7: Kết quả Pivot Table thống kê tổng lợi nhuận theo loại sản phẩm**

* **Manual SSAS**



**Hình 4.8: Kết quả Manual SSAS thống kê tổng lợi nhuận theo loại sản phẩm**

### 2.3. Thống kê số lượng bán (Quantity) theo quốc gia (Country) - Roll up

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.9: Kết quả truy vấn MDX thống kê số lượng bán theo quốc gia**

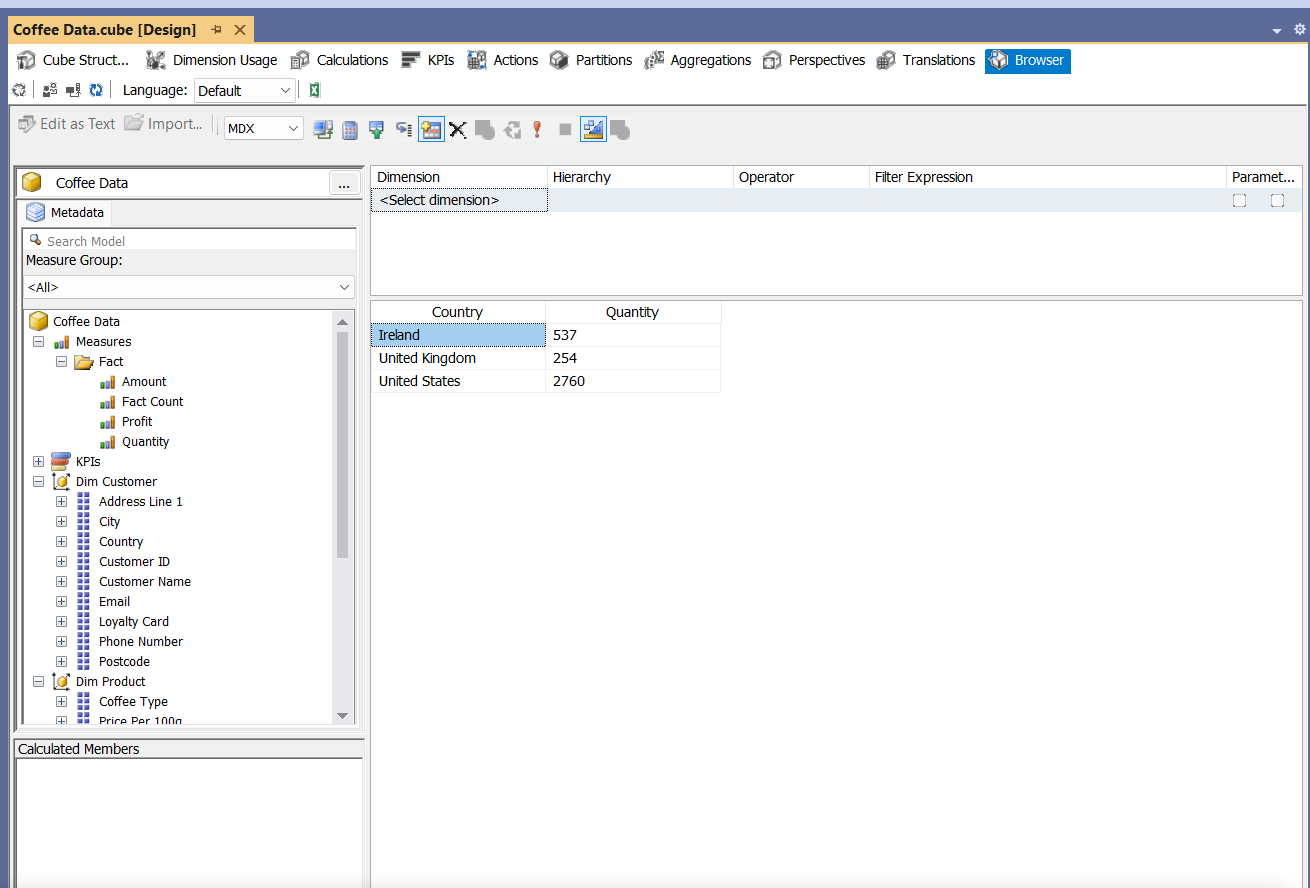
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.10: Kết quả Pivot table thống kê số lượng bán theo quốc gia**

* **Manual SSAS**



**Hình 4.11: Kết quả Manual SSAS thống kê số lượng bán theo quốc gia**

### 2.4. Thống kê lợi nhuận (Profit) theo thành phố (City) trong năm 2022 - Drill down

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.12: Kết quả truy vấn MDX thống kê lợi nhuận theo thành phố năm 2022**

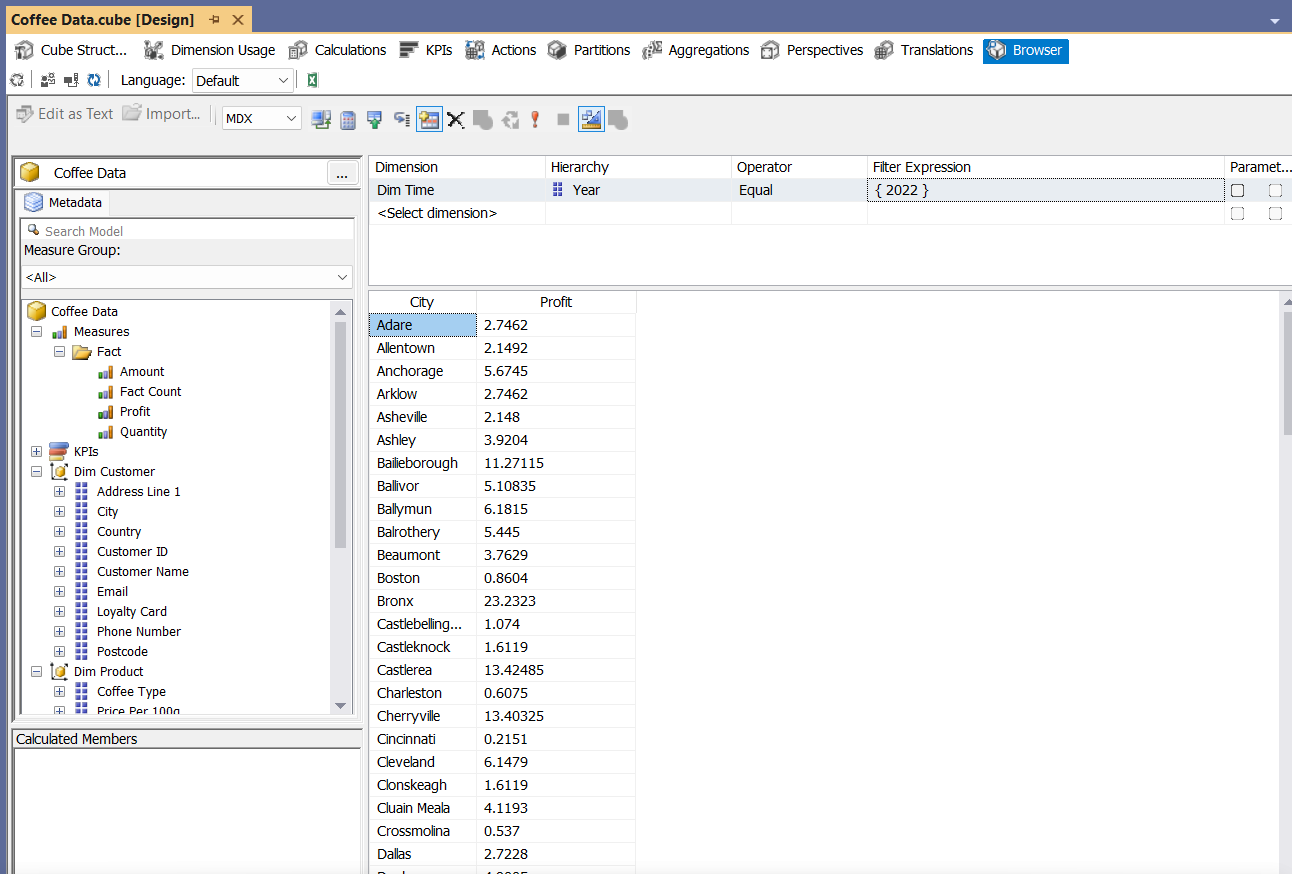
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.13: Kết quả Pivot table thống kê lợi nhuận theo thành phố năm 2022**

* **Manual SSAS**



**Hình 4.14: Kết quả Manual SSAS thống kê lợi nhuận theo thành phố năm 2022**

### 2.5. Thống kê doanh thu (Amount) theo phân loại khách hàng Loyalty - Slice and Dice

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.15: Kết quả truy vấn MDX thống kê doanh thu theo phân loại khách hàng**

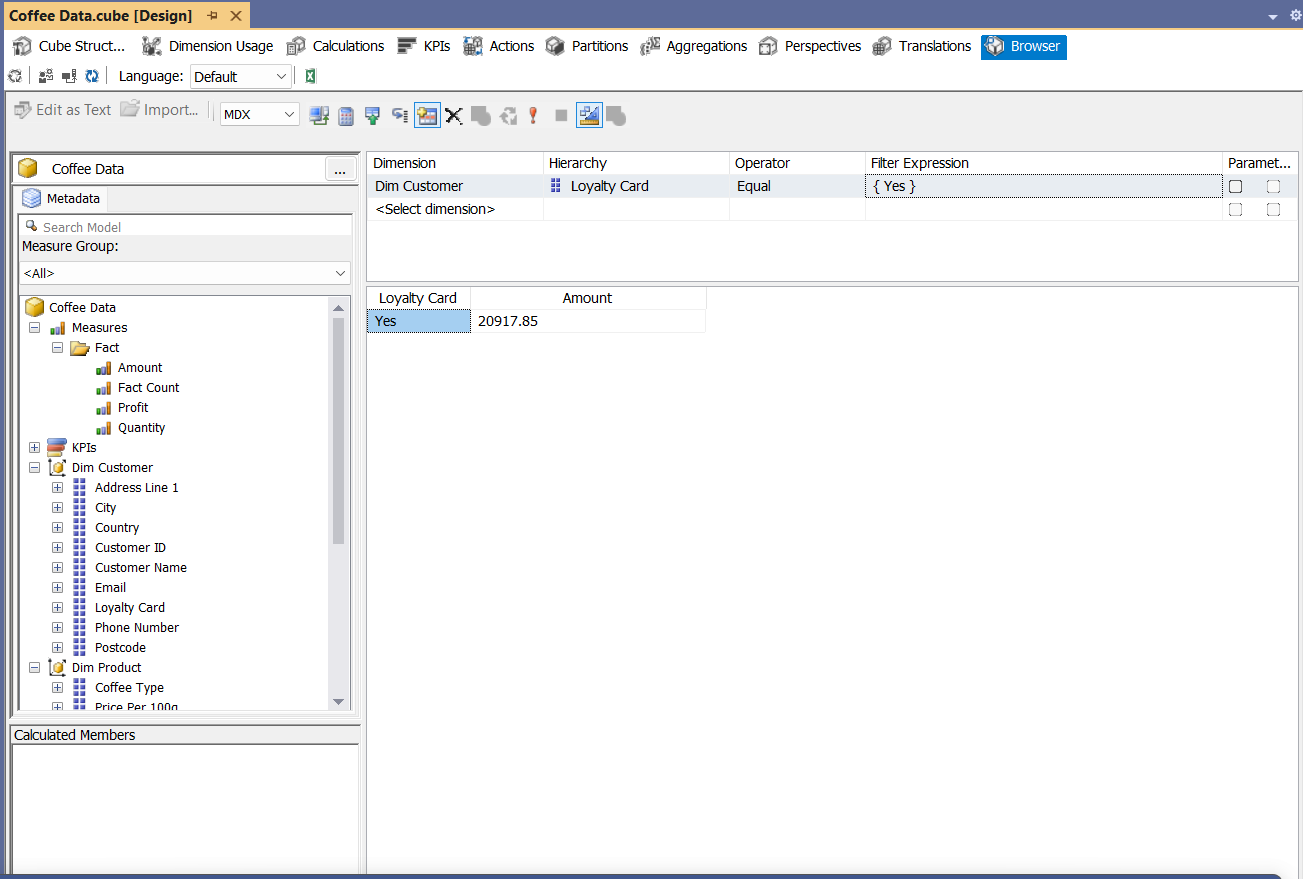
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.16: Kết quả Pivot table thống kê doanh thu theo phân loại khách hàng**

* **Manual SSAS**



**Hình 4.17: Kết quả Manual SSAS thống kê doanh thu theo phân loại khách hàng**

### 2.6. Thống kê doanh thu (Amount) theo từng tháng trong năm - Drill down

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.18: Kết quả truy vấn MDX thống kê doanh thu theo từng tháng trong năm**

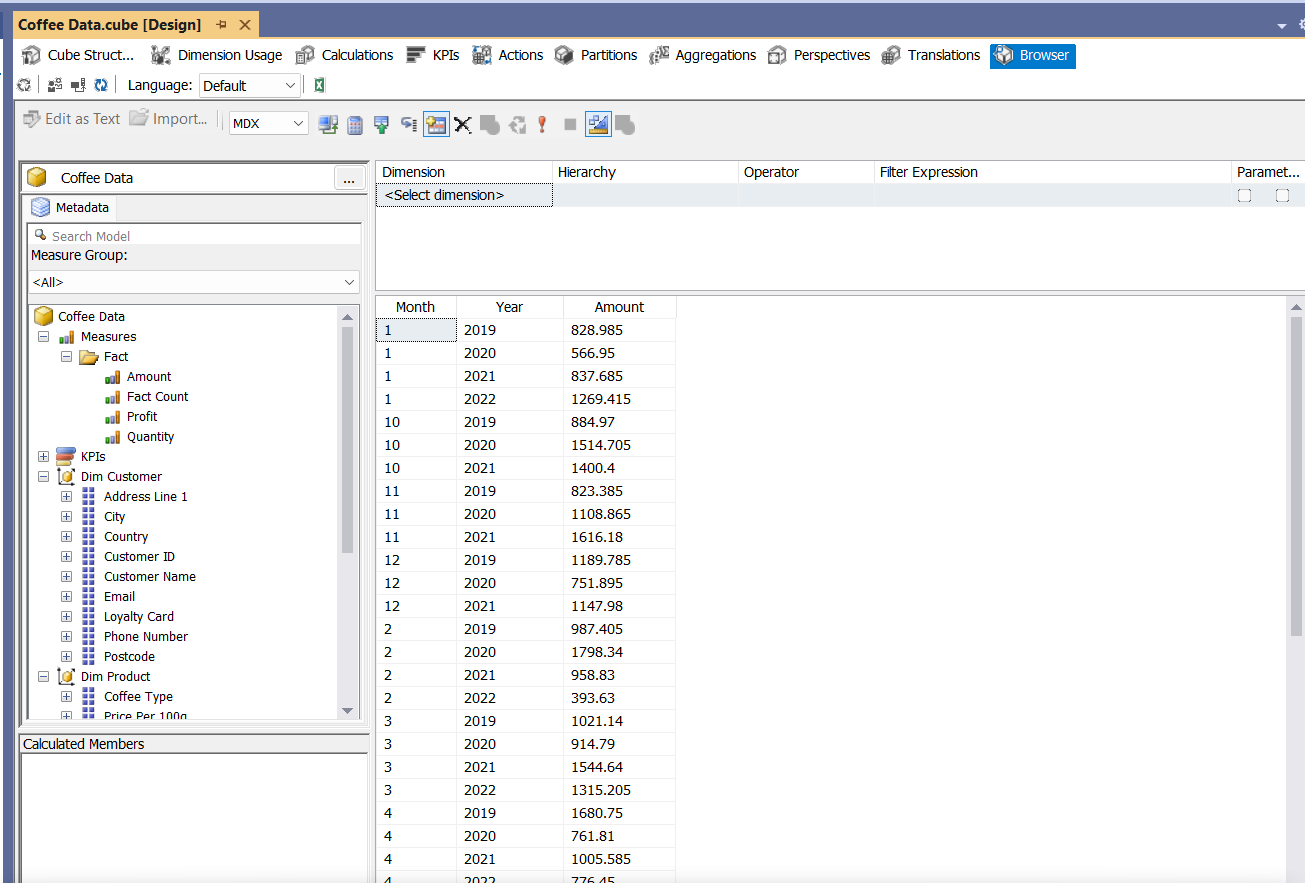
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.19: Kết quả Pivot table thống kê doanh thu theo từng tháng trong năm**

* **Manual SSAS**



**Hình 4.20: Kết quả Manual SSAS thống kê doanh thu theo từng tháng trong năm**

### Thống kê 5 Quốc gia (Country) có doanh thu (Amount) cao nhất - Pivot

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.21: Kết quả truy vấn MDX thống kê 5 Quốc gia có doanh thu cao nhất**

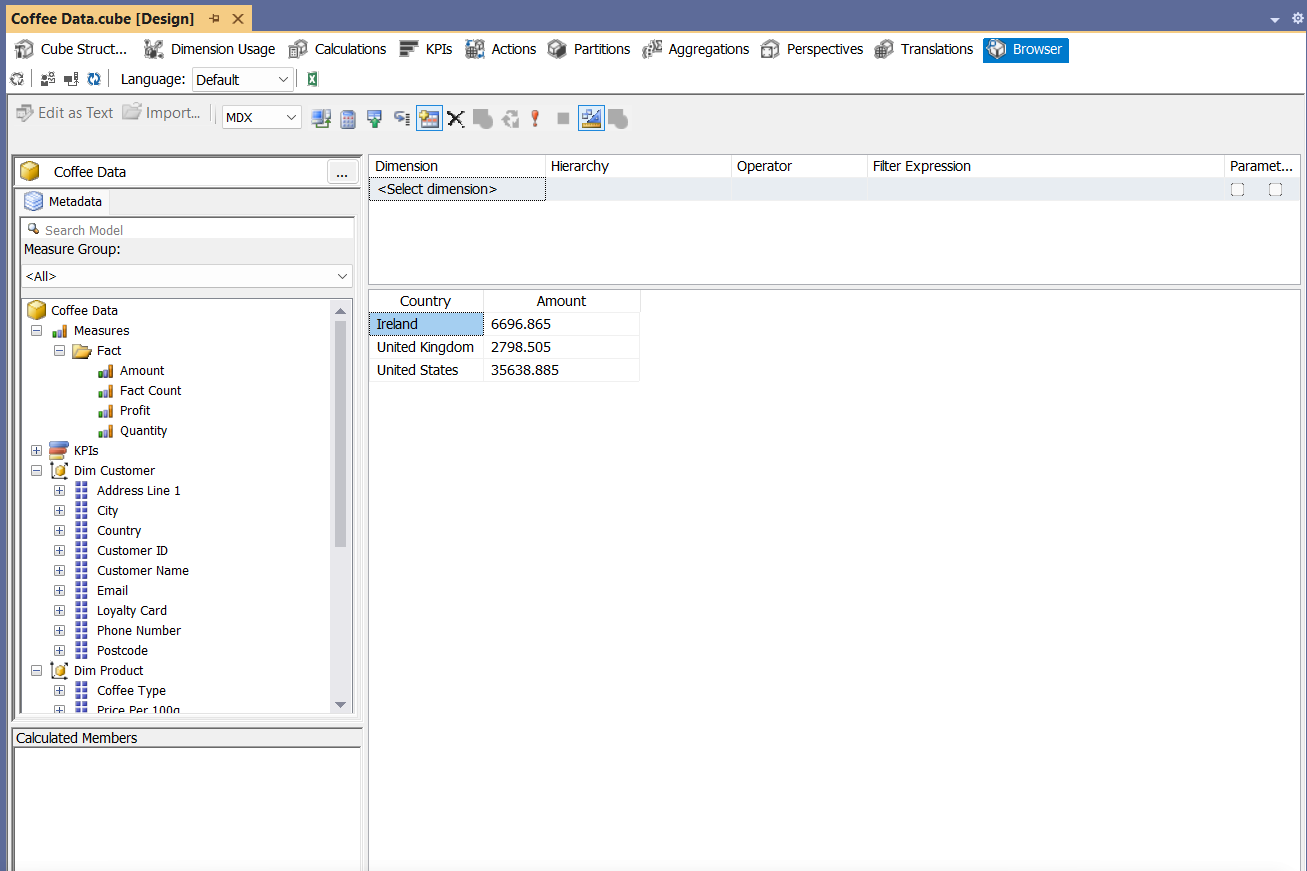
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.22: Kết quả Pivot table thống kê 5 Quốc gia có doanh thu cao nhất**

* **Manual SSAS**



**Hình 4.23: Kết quả Manual SSAS thống kê 5 Quốc gia có doanh thu cao nhất**

### 2.8. Thống kê số lượng sản phẩm bán ra (Quantity) theo các ngày trong tuần (Day of Week) - Slice and Dice

* **MDX**A screenshot of a computer program

  Description automatically generated

**Hình 4.24: Kết quả MDX thống kê số lượng sản phẩm bán ra theo ngày trong tuần**

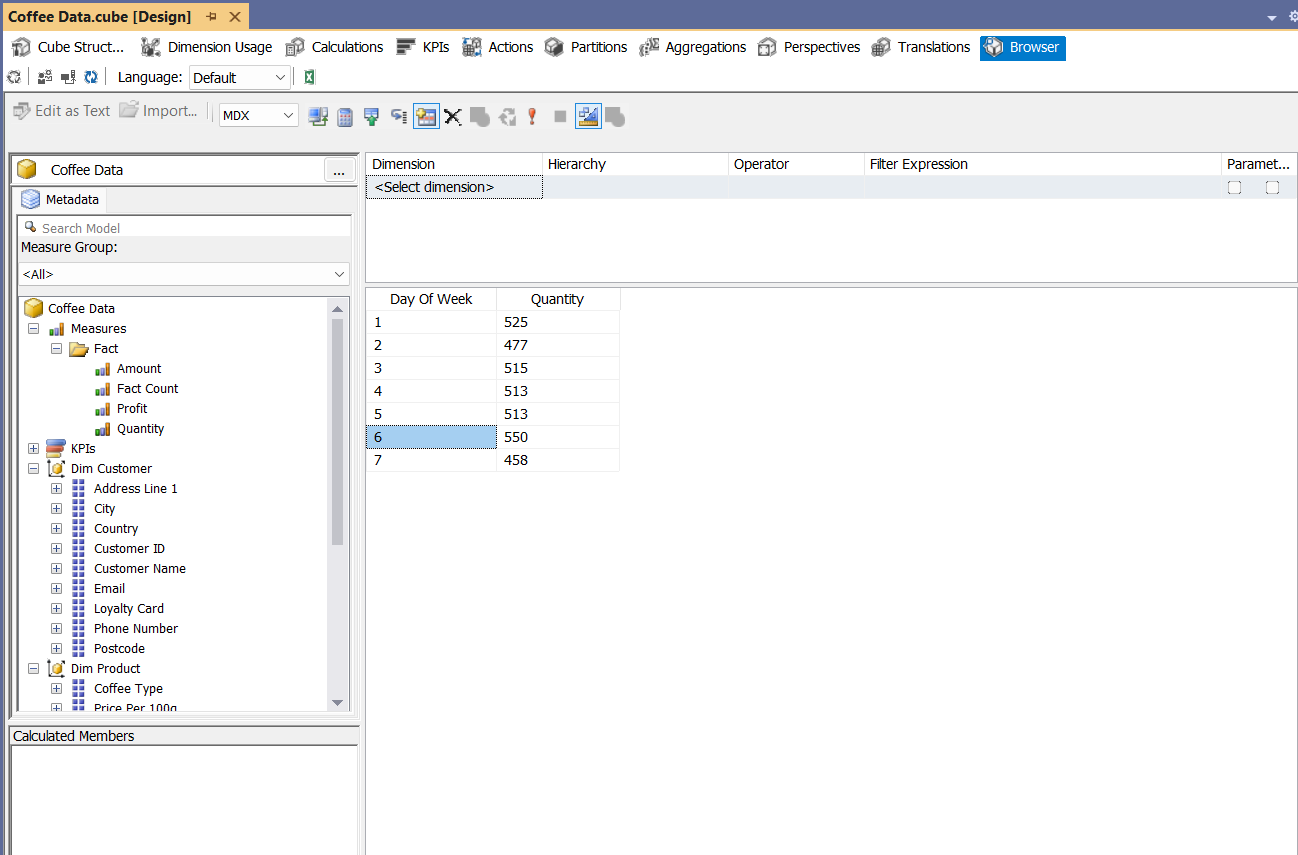
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.25: Kết quả Pivot table thống kê số lượng sản phẩm bán ra theo các ngày trong tuần**

* **Manual SSAS**



**Hình 4.26: Kết quả Manual SSAS thống kê số lượng sản phẩm bán ra theo các ngày trong tuần**

### 2.9. Thống kê sản lượng loại cà phê (Coffee Type) và cách rang (Roast Type) được khách hàng yêu thích - Slice and Dice

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.27: Kết quả MDX thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích**

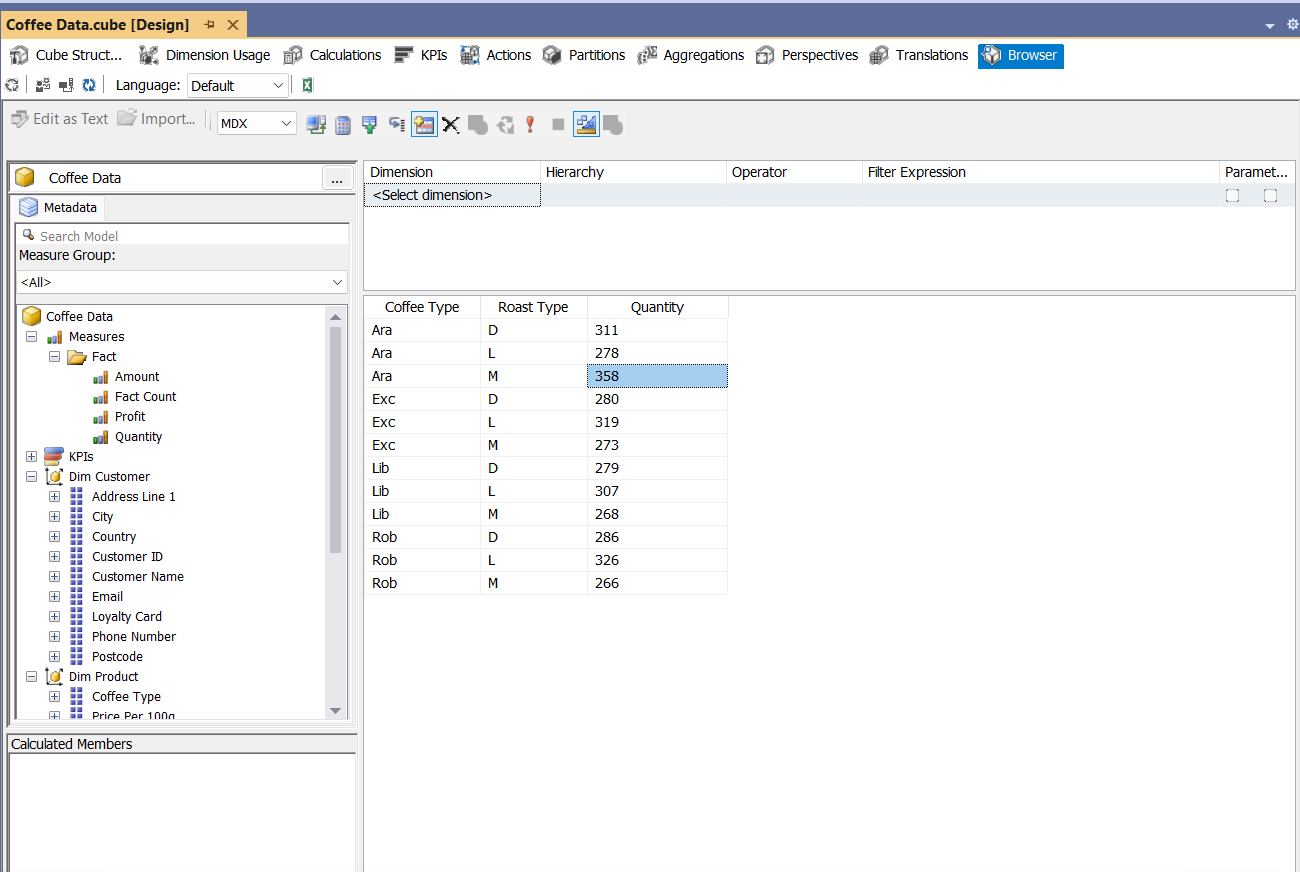
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.28: Kết quả Pivot table thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích**

* **Manual SSAS**



**Hình 4.29: Kết quả Manual SSAS thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích**

### 2.10. Thống kê sản phẩm có doanh thu (Profit) cao nhất - Pivot

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.30: Kết quả MDX thống kê sản phẩm có doanh thu cao nhất**

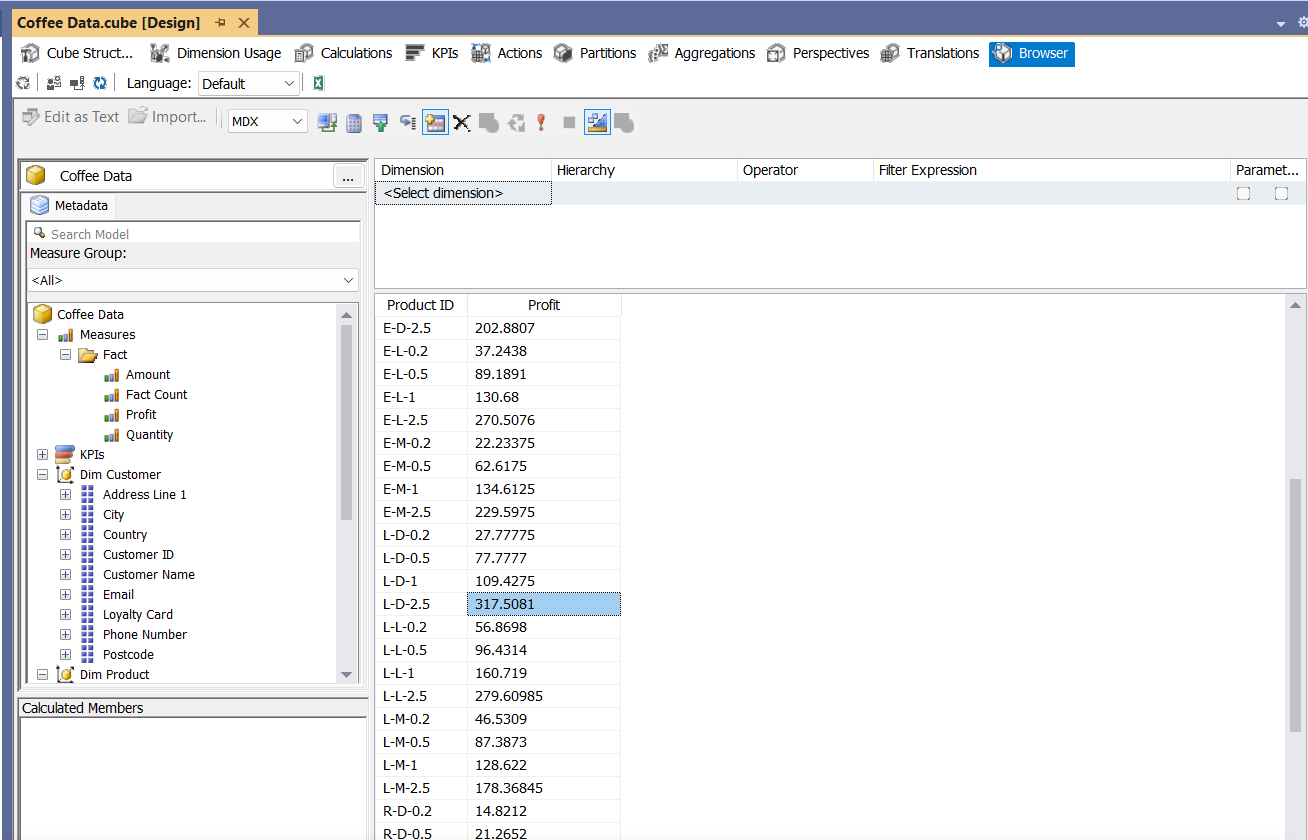
* **Pivot table trên MS Excel**

A screenshot of a computer

Description automatically generated

**Hình 4.31: Kết quả Pivot table thống kê sản phẩm có doanh thu cao nhất**

* **Manual SSAS**



**Hình 4.32: Kết quả Manual SSAS thống kê sản phẩm có doanh thu cao nhất**

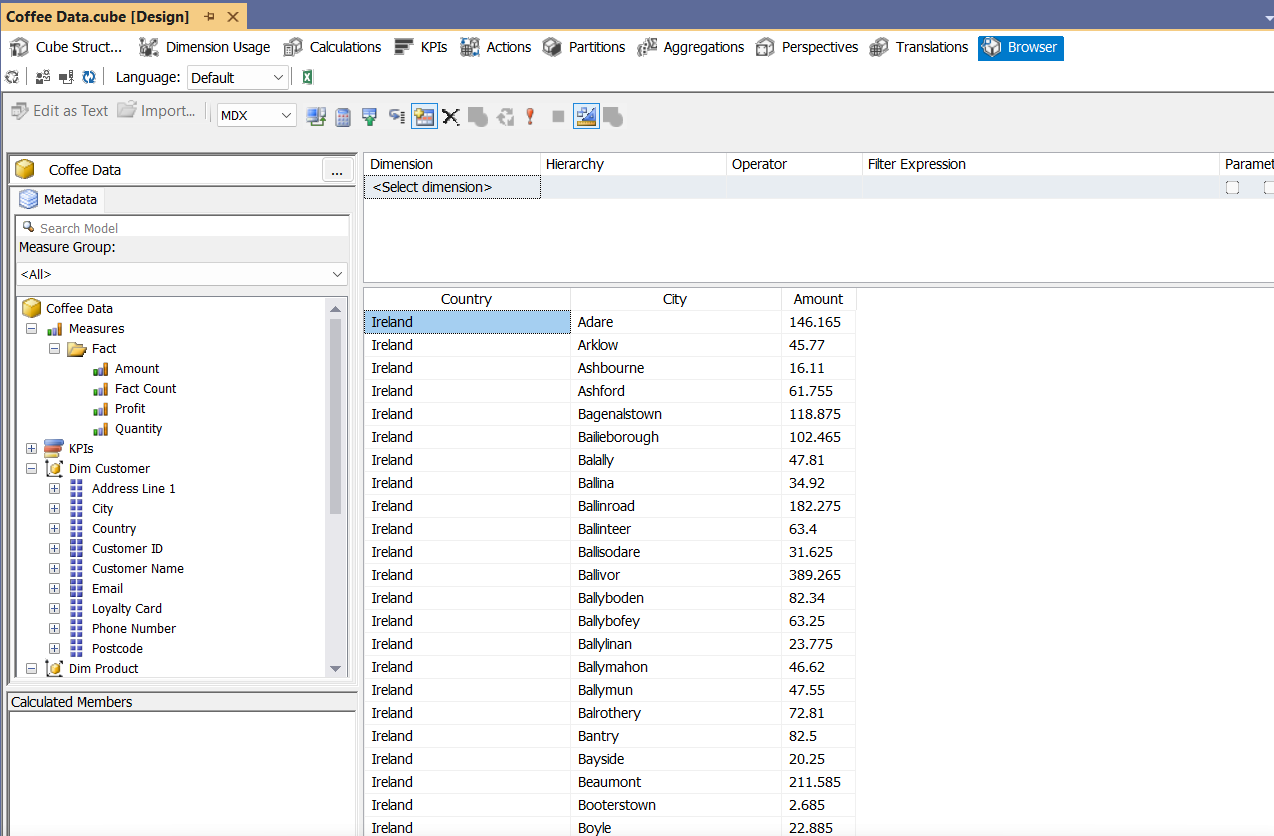
### 2.11. Thống kê doanh thu mỗi thành phố (City) của mỗi quốc gia (Country) - Drill down

* **MDX**A screenshot of a computer

  Description automatically generated

**Hình 4.33: Kết quả MDX thống kê doanh thu mỗi thành phố của quốc gia**

* **Manual SSAS**

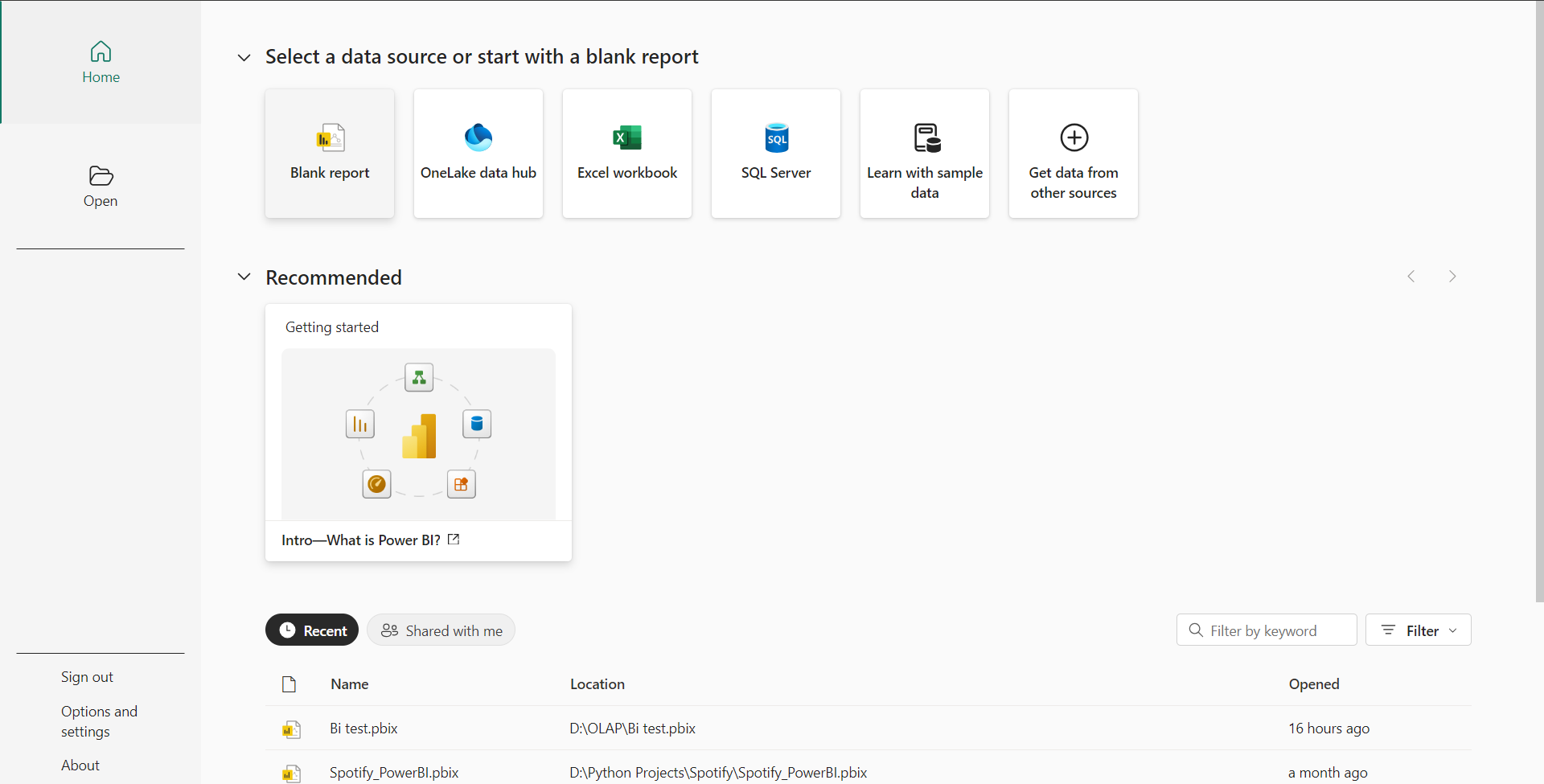


**Hình 4.34: Kết quả Manual SSASthống kê doanh thu mỗi thành phố của quốc gia**

# Chương 5. QUÁ TRÌNH REPORT POWER BI (SSRS)

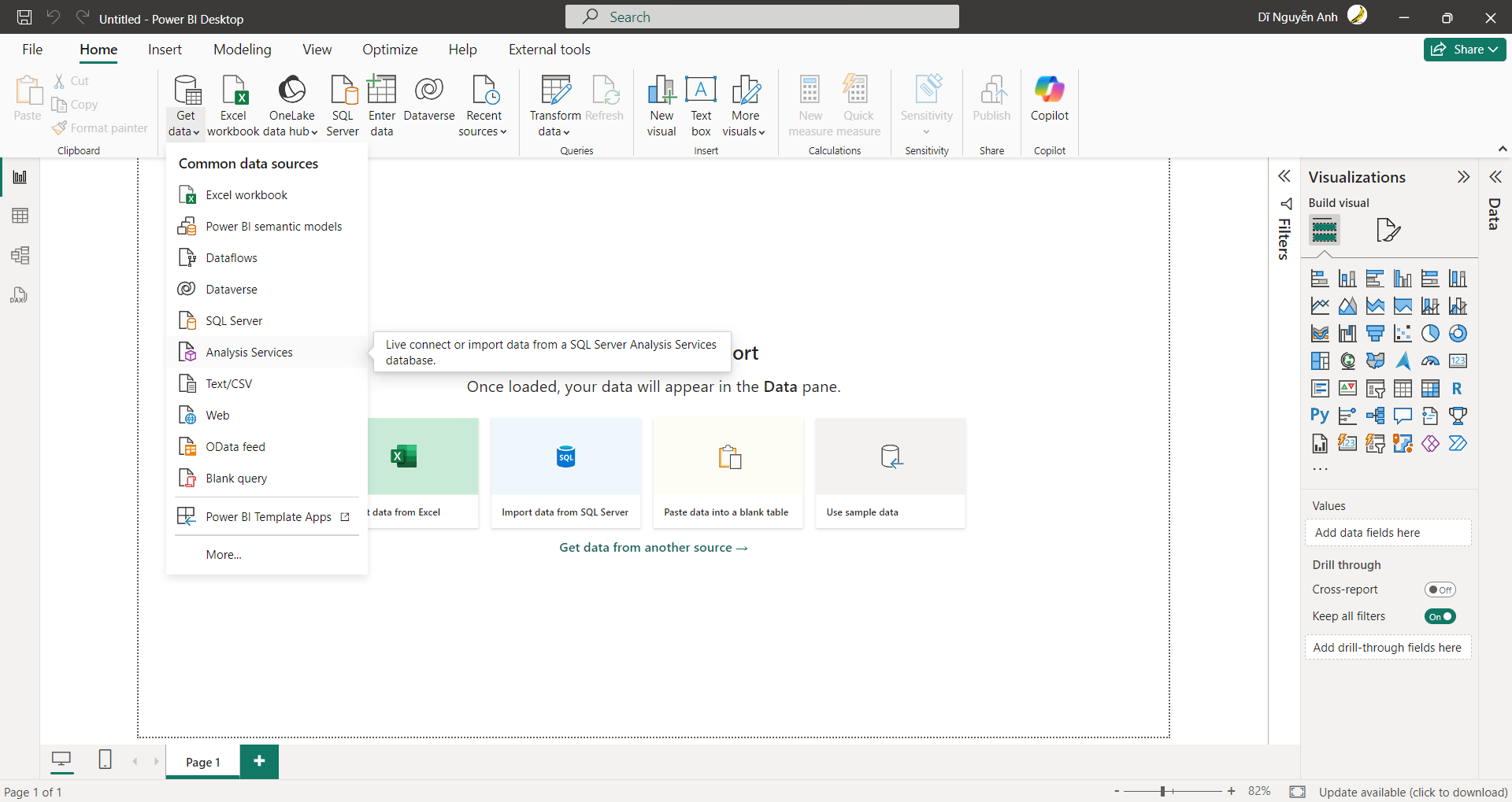
## 1. Đổ dữ liệu từ SSAS vào Power BI

Bước 1: Vào phần mềm **Power BI** và bấm vào mục **Blank Report**

****

**Hình 5.1: Giao diện phần mềm Power BI**

**Bước 2**: Bấm vào mục **Get data** -> chọn **Analysis Services**

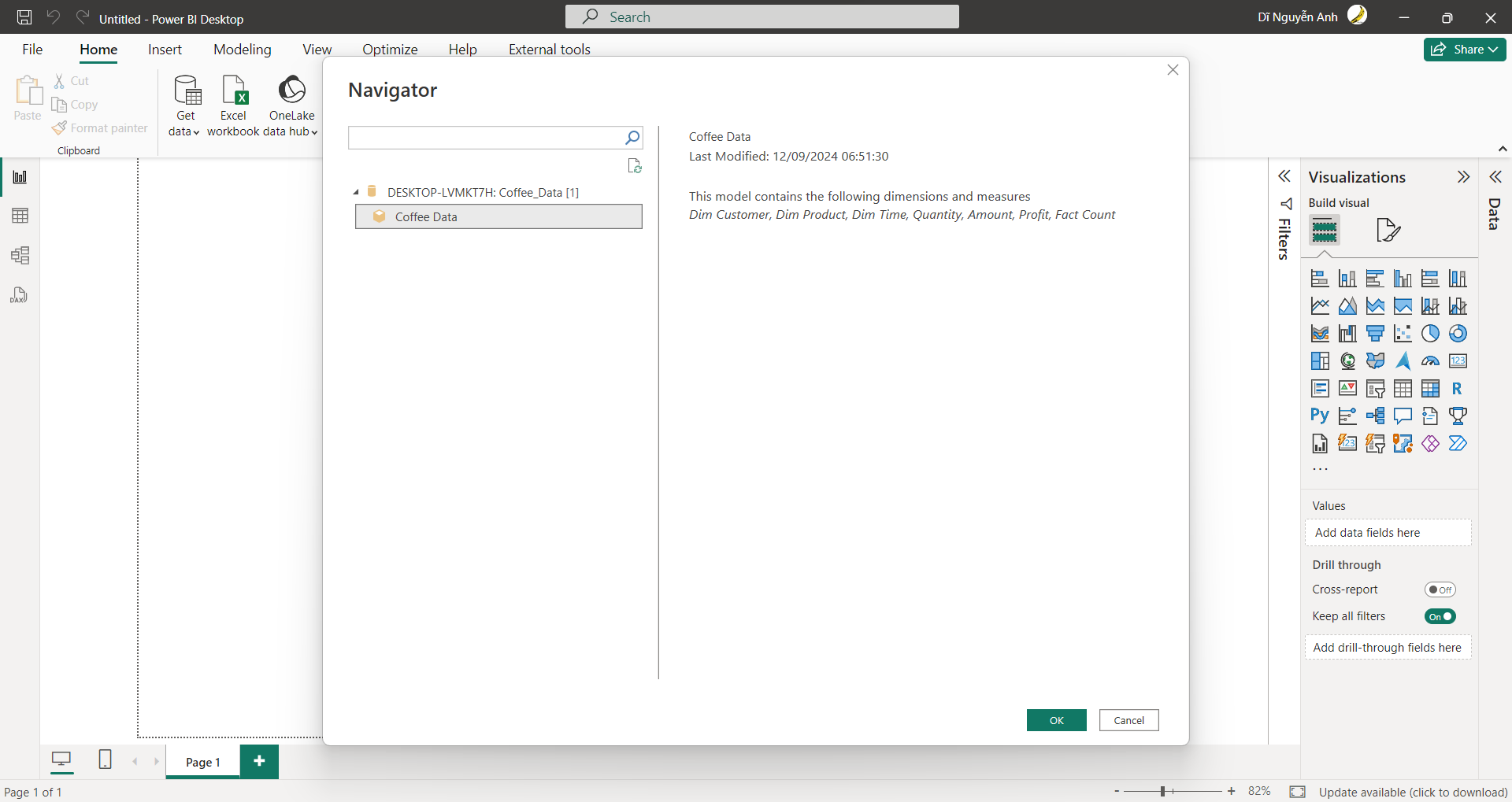
****

**Hình 5.2: Giao diện chọn công cụ Analysis Services trong Power BI**

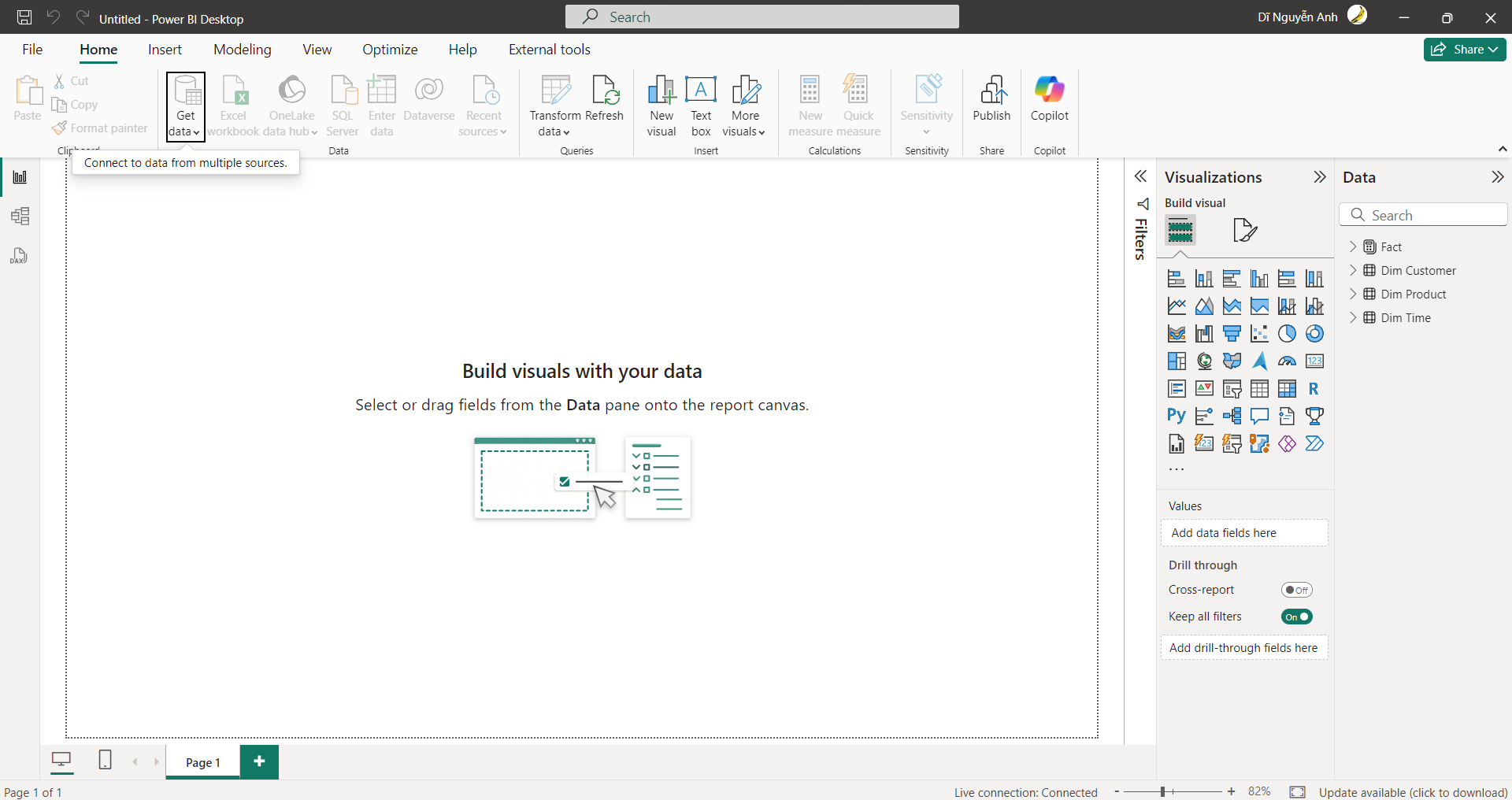
**Bước 3**: Sau đó ta thấy được một pop-up xuất hiện, tại đây ta sẽ điền **Server** theo tên server của SQL, và tên database tương ứng -> và nhấn **OK**

**Hình 5.3: Giao diện chọn Server từ Microsoft SQL Server**

**Bước 4**: Bảng Navigator sẽ xuất hiện, Chọn Cube chứa thông tin để lấy dữ liệu → Nhấn OK

****

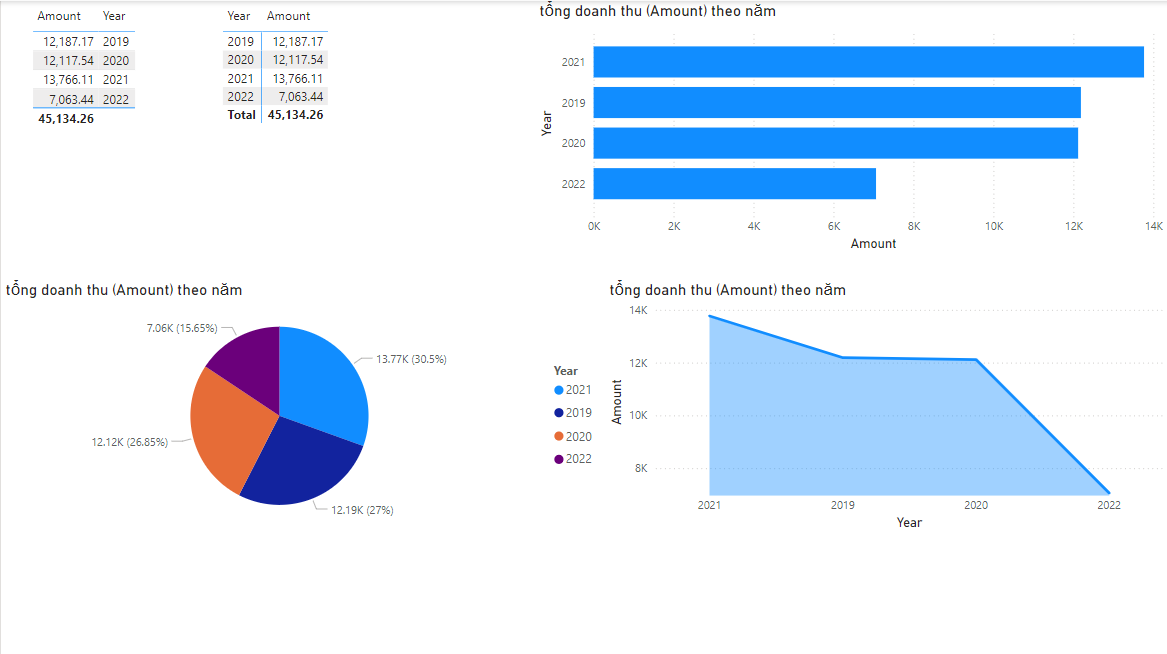
**Hình 5.4: Giao diện chọn Cube để lấy dữ liệu từ quá trình SSAS**

**Bước 5**: Tại bên phải, mục **Data**, ta thấy được **Fact** cùng với các bảng **Dim** khác, tức dữ liệu được đổ về từ **Analysis Service** đã hoàn tất****

**Hình 5.5: Giao diện đổ dữ liệu từ Analysis Service hoàn tất**

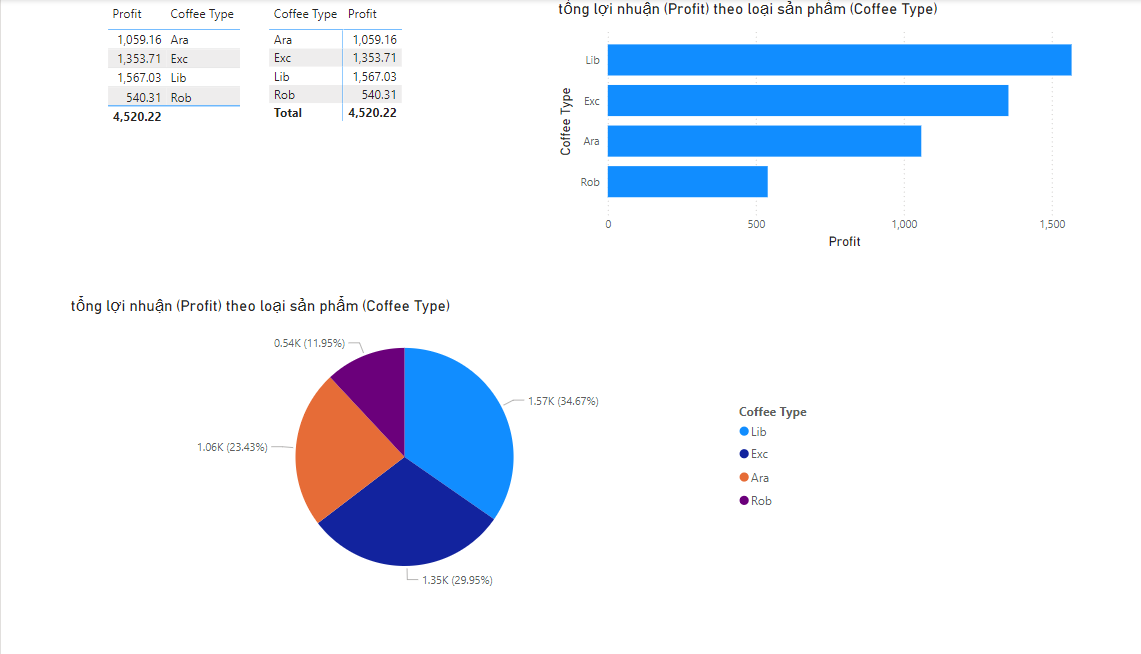
## 2. Thực hiện SSRS

### 2.1. Thống kê tổng số doanh thu (Amount) theo năm (Year)



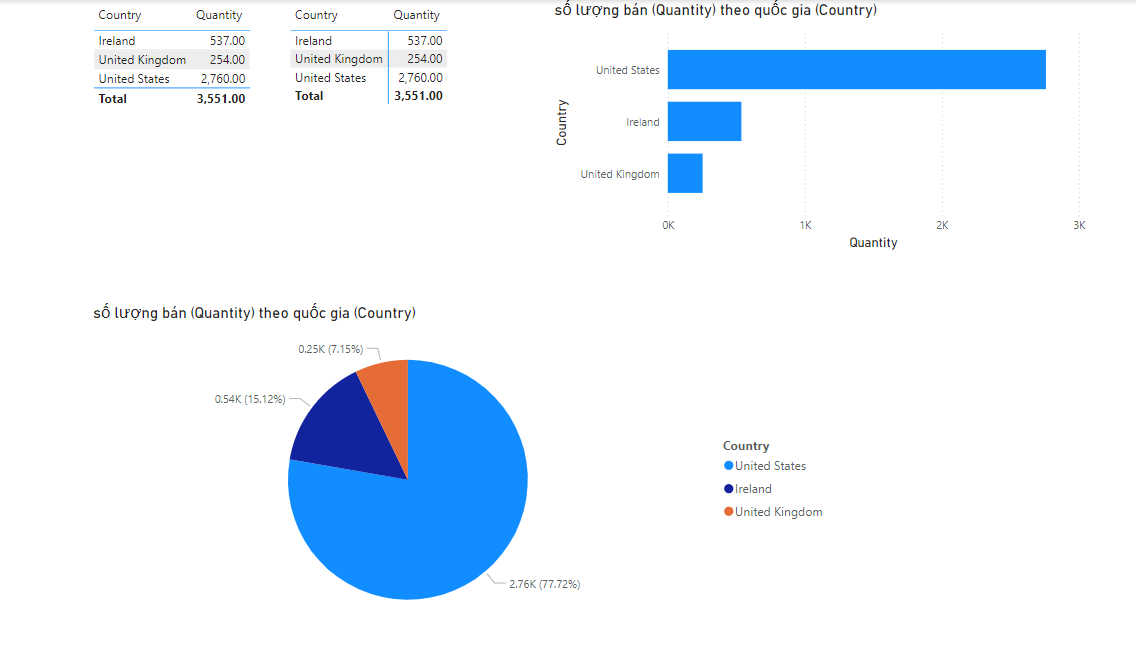
**Hình 5.6: Giao diện thống kê phân tích tổng doanh thu theo năm**

### 2.2. Thống kê tổng số lợi nhuận (Profit) theo loại sản phẩm (Coffee Type)



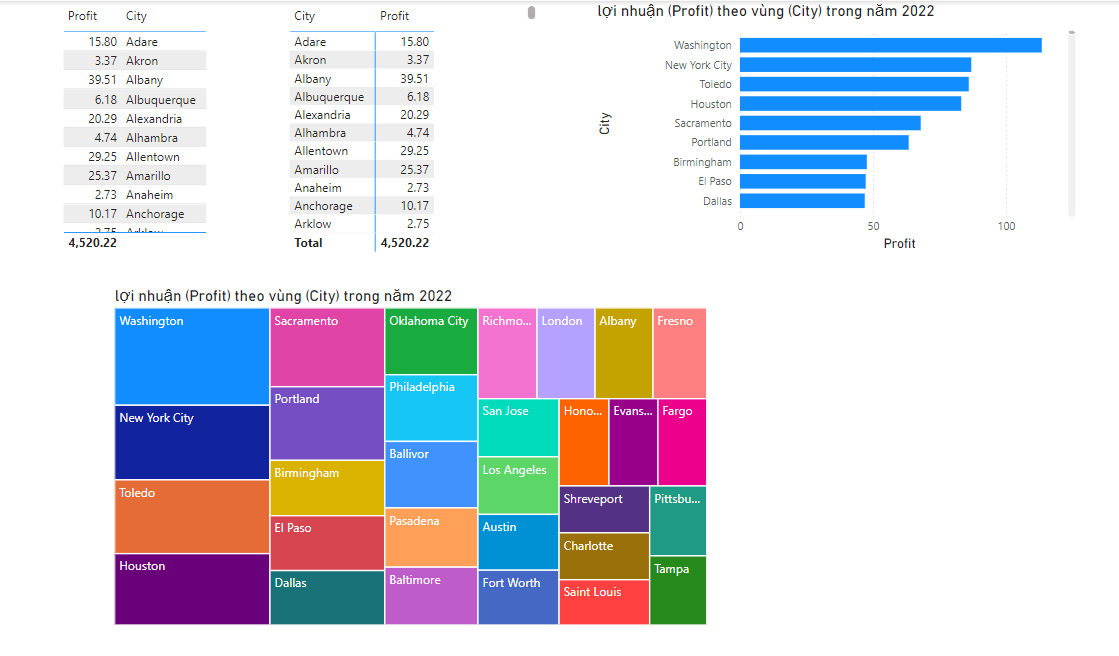
**Hình 5.7: Giao diện thống kê tổng lợi nhuận theo loại sản phẩm**

### 2.3. Thống kê số lượng bán (Quantity) theo quốc gia (Country)



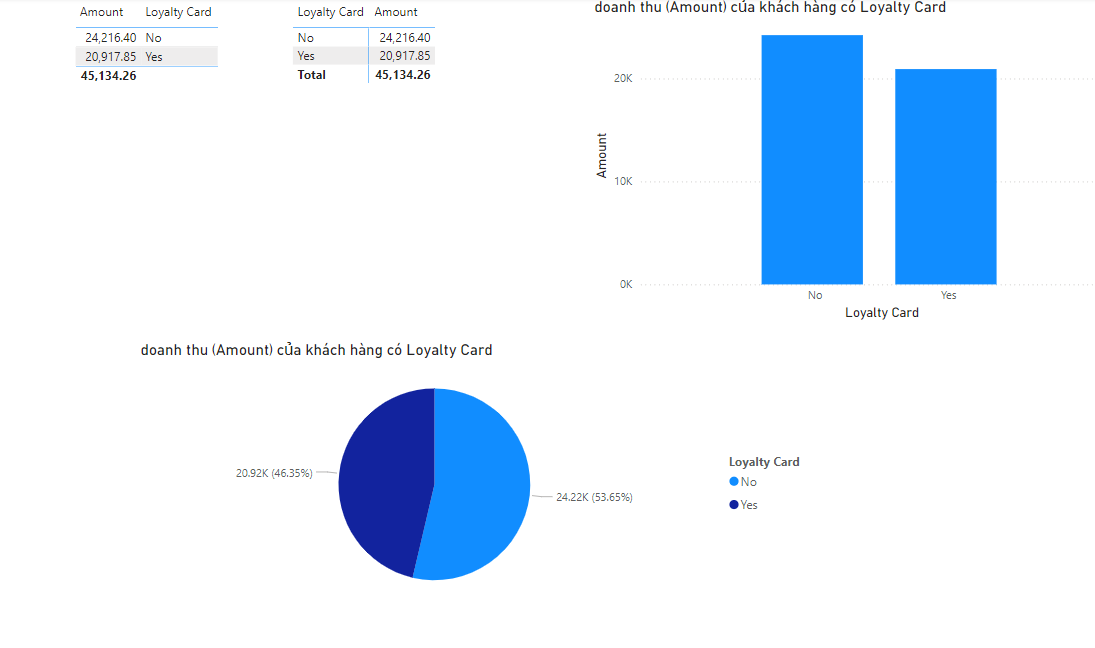
**Hình 5.8: Giao diện thống kê số lượng bán theo quốc gia**

### 2.4. Thống kê lợi nhuận (Profit) theo vùng (City) trong năm 2022



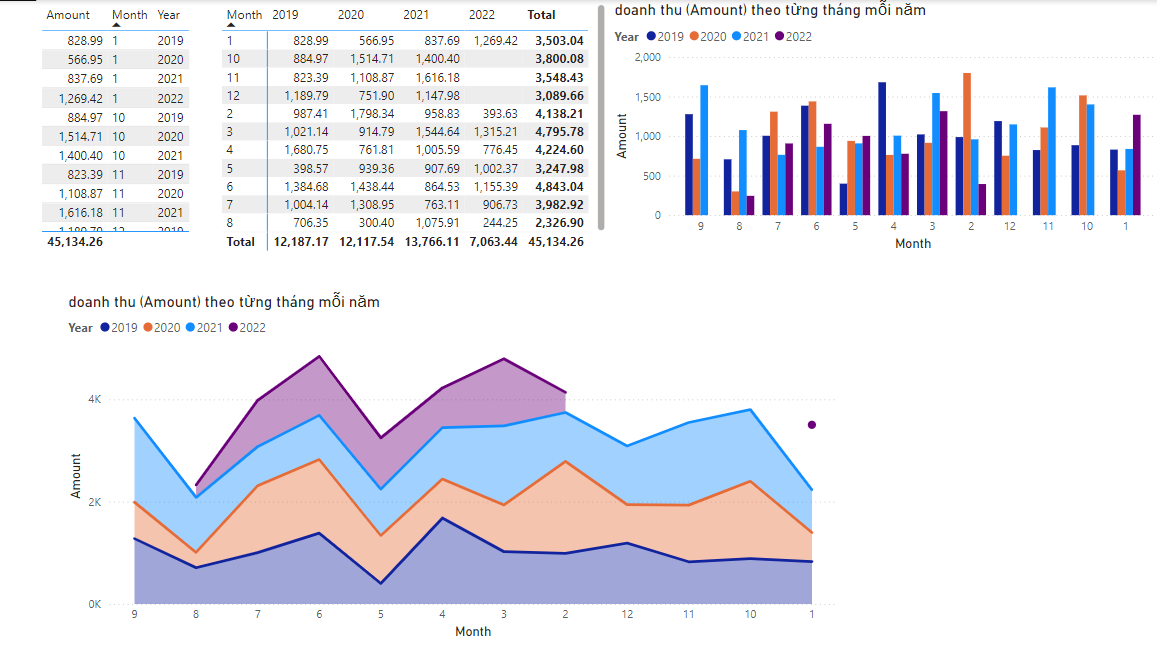
**Hình 5.9: Giao diện thống kê lợi nhuận theo thành phố năm 2022**

### 2.5. Thống kê tổng số doanh thu (Amount) của khách hàng có Loyalty Card



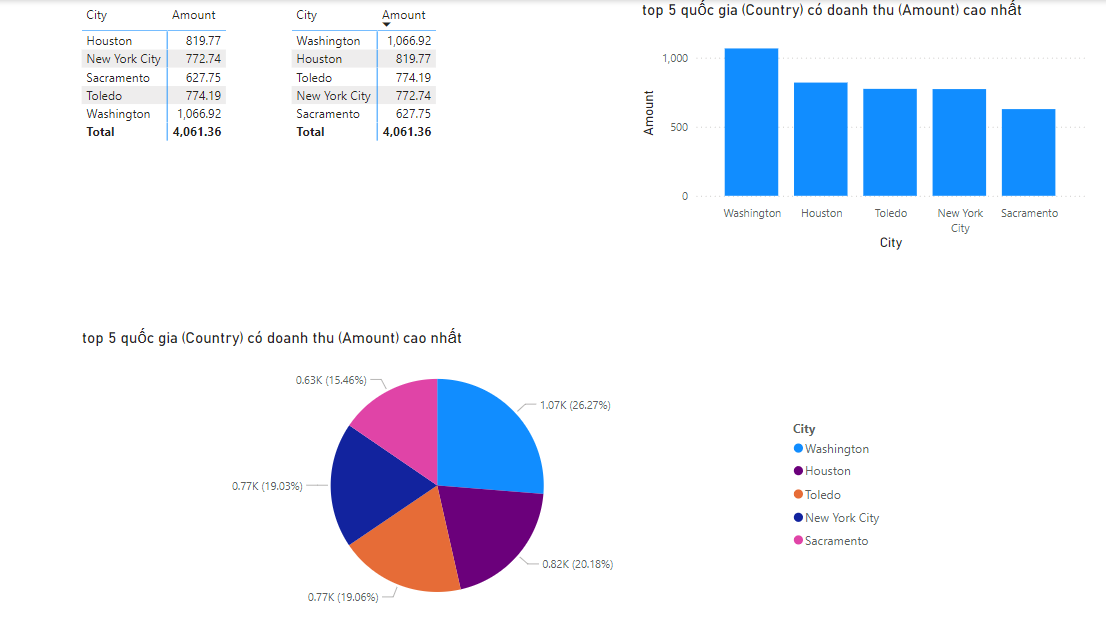
**Hình 5.10: Giao diện thống kê doanh thu theo phân loại khách hàng**

### 2.6. Thống kê doanh thu (Amount) theo từng tháng mỗi năm



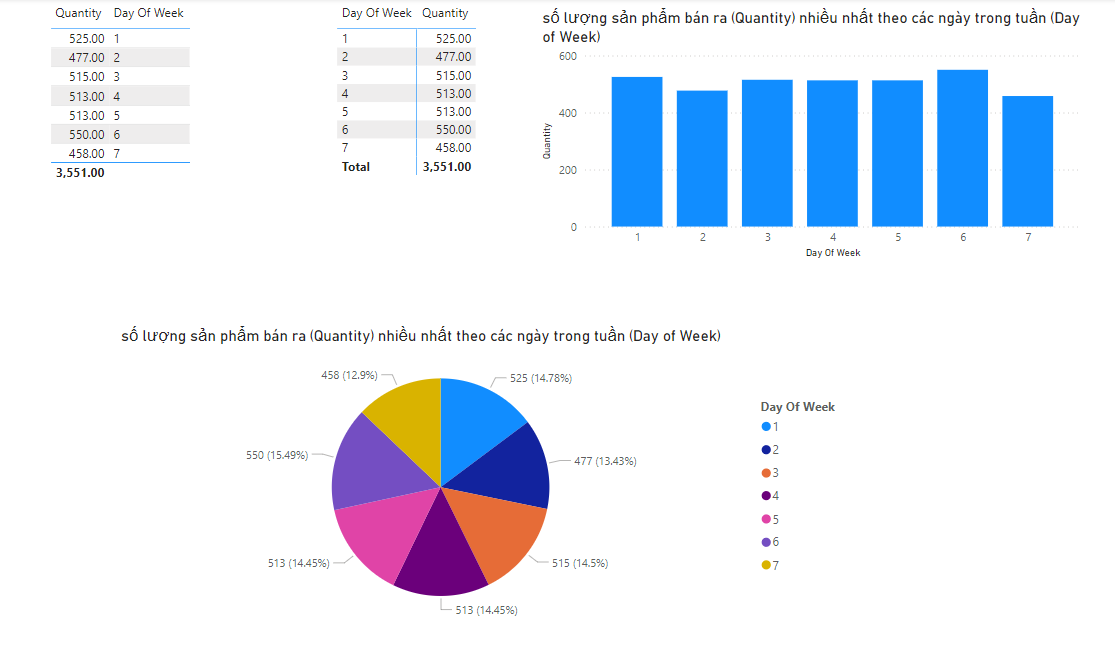
**Hình 5.11: Giao diện thống kê doanh thu theo từng tháng trong năm**

### 2.7. Thống kê top 5 quốc gia (Country) có doanh thu (Amount) cao nhất



**Hình 5.12: Giao diện thống kê 5 Quốc gia có doanh thu cao nhất**

### 2.8. số lượng sản phẩm bán ra (Quantity) nhiều nhất theo các ngày trong tuần (Day of Week)



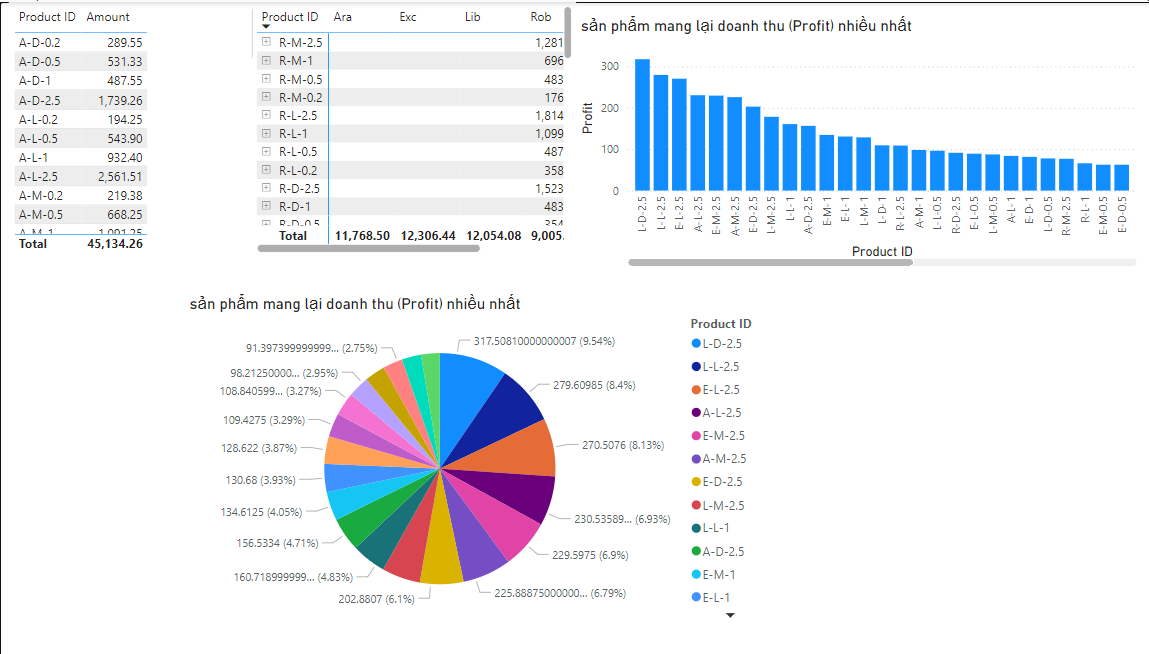
**Hình 5.13: Giao diện thống kê**

### 2.9. Thống kê khách hàng thích loại cà phê (Coffee Type) và cách rang (Roast Type) nào



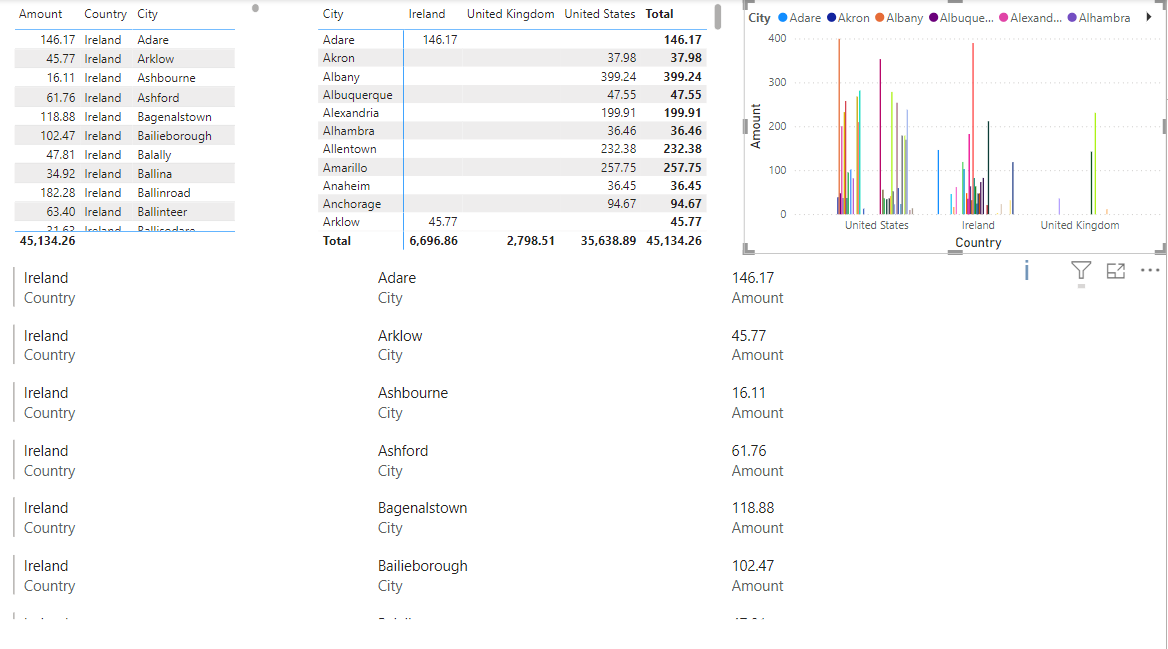
**Hình 5.14: Giao diện thống kê sản lượng loại cà phê và cách rang được khách hàng yêu thích**

### ***2.10. Thống kê sản phẩm mang lại doanh thu (Profit) nhiều nhất***



**Hình 5.15: Giao diện thống kê sản phẩm có doanh thu cao nhất**

### 2.11. Thống kê doanh thu của từng quốc gia (Country) cho từng thành phố (City)



**Hình 5.16: Giao diện thống kê doanh thu mỗi thành phố của quốc gia**

# Chương 6. QUÁ TRÌNH KHAI PHÁ DỮ LIỆU

## 1. Giới thiệu

Trong bối cảnh hiện nay, dự đoán doanh thu là một phần quan trọng của việc quản lý kinh doanh. Bộ dữ liệu Coffee Bean Sales Raw Dataset, một bộ dữ liệu doanh số bán cà phê tại Mỹ được công bố trên trang Kaggle.com, cung cấp một cơ hội tuyệt vời để áp dụng các thuật toán dự đoán nhằm nâng cao hiệu quả kinh doanh. Chúng tôi đã triển khai bốn thuật toán dự đoán khác nhau để dự đoán doanh thu trong 1 năm kế tiếp dựa trên bộ dữ liệu này. Các thuật toán được sử dụng bao gồm Linear Regression, Random Forest (RF), Long Short-Term Memory (LSTM), và Moving Averages.

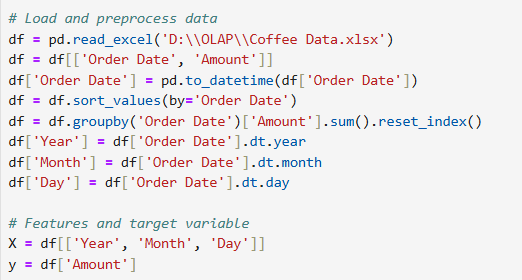
* Linear Regression: Là một trong những phương pháp đơn giản nhất nhưng vẫn hiệu quả. Hồi quy tuyến tính xác định mối quan hệ tuyến tính giữa doanh thu và các yếu tố khác, từ đó dự đoán doanh thu trong tương lai. Phương pháp này phù hợp khi mối quan hệ giữa các biến là tuyến tính.
* Random Forest - RF: Một phương pháp học máy mạnh mẽ dựa trên tập hợp các cây quyết định. Bằng cách xây dựng nhiều cây quyết định và tổng hợp kết quả, RF có thể nắm bắt các mối quan hệ phức tạp trong dữ liệu, kể cả phi tuyến tính, và cải thiện độ chính xác của dự đoán so với cây quyết định đơn lẻ. RF cũng ít bị ảnh hưởng bởi overfitting.
* Long Short-Term Memory - LSTM: Một loại mạng nơ-ron hồi quy (RNN) đặc biệt hiệu quả với dữ liệu chuỗi thời gian. LSTM có khả năng ghi nhớ thông tin trong khoảng thời gian dài, giúp nắm bắt được cả xu hướng dài hạn và biến động ngắn hạn trong dữ liệu doanh thu, từ đó đưa ra dự đoán chính xác hơn. Phương pháp này phù hợp với dữ liệu có tính chu kỳ hoặc xu hướng theo thời gian.
* Moving Averages: Làm mịn dữ liệu chuỗi thời gian bằng cách tính trung bình của một tập hợp các điểm dữ liệu liên tiếp. Các biến thể bao gồm trung bình di động đơn giản (SMA), trung bình di động có trọng số (WMA) và trung bình di động hàm mũ (EMA). Hữu ích để xác định xu hướng chung và dự đoán ngắn hạn, nhưng có thể bị trễ và không hiệu quả cho dự đoán dài hạn.

## 2. Các bước thực hiện theo từng thuật toán

## 2.1. Random Forest

Tiến hành Tiến hành nhập dữ liệu file csv, đổi định dạng cột ‘Order Date’ thành timestamp, sau đó sắp xếp DataFrame theo cột ‘Order Date’. Nhóm ‘Amount’ theo ngày. Chia ‘Order Date’ thành ‘Year’, ‘Month’, ‘Day’.

Biến ‘Year’, ‘Month’, ‘Day’ là biến giải thích, biến ‘Amount’ là biến phẩn hồi



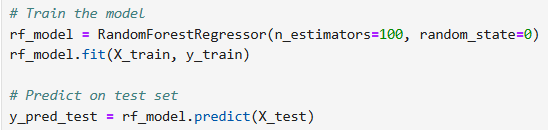
**Hình 6.1: Import và sắp xếp dữ liệu**

Chia tập train, test theo tỷ lệ 70:30



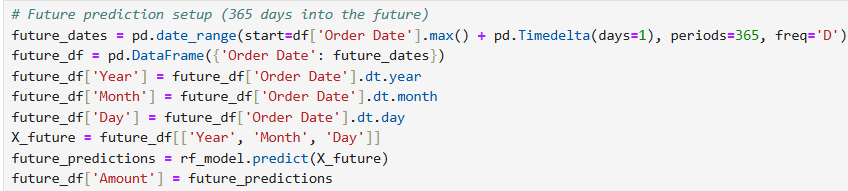
**Hình 6.2: Chia tập train, test theo tỉ lệ 7:3**

Thực hiện train model với mô hình Random Forest với tham số đầu vào là n\_estimators=100, random\_state=0



**Hình 6.3: Truyền các tham số đầu vào cho mô hình Random Forest**

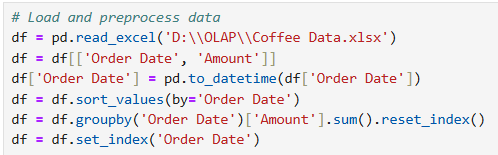
Dự đoán 1 năm sau bằng các tạo các ngày tính từ thời điểm gần nhất của data, thực hiện dự đoán ‘Amount’ theo ngày



**Hình 6.4: Dự đoán doanh thu 1 năm sau từ thời điểm gần nhất**

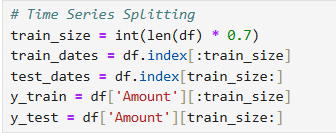
## 2.2. Moving Averages

Tiến hành Tiến hành nhập dữ liệu file csv, đổi định dạng cột ‘Order Date’ thành timestamp, sau đó sắp xếp DataFrame theo cột ‘Order Date’. Nhóm ‘Amount’ theo ‘Order Date’.



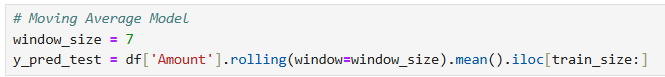
**Hình 6.5: Import và sắp xếp dữ liệu**

Chia tập train test theo tỷ lệ 70:30



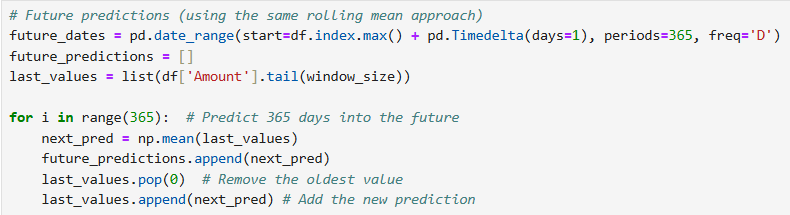
**Hình 6.6: Chia tập train, test theo tỉ lệ 7:3**

Dự đoán tập test bằng thuật toán Moving Averages với giá trị window size là



**Hình 6.7: Chọn giá trị window size và dự đoán tập test**

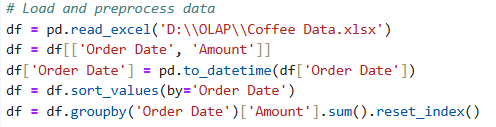
Dự đoán 1 năm sau bằng các tạo các ngày tính từ thời điểm gần nhất của data, thực hiện dự đoán ‘Amount’ theo ngày



**Hình 6.8: Dự đoán doanh thu 1 năm sau từ thời điểm gần nhất**

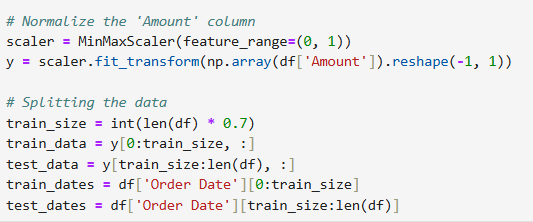
## 2.3. LSTM

Tiến hành Tiến hành nhập dữ liệu file csv, đổi định dạng cột ‘Order Date’ thành timestamp, sau đó sắp xếp DataFrame theo cột ‘Order Date’. Nhóm ‘Amount’ theo ngày.



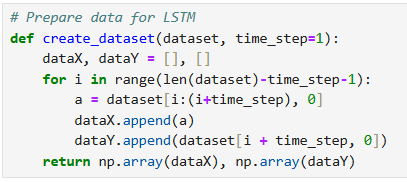
**Hình 6.9: Import và nhóm cột Amount theo ngày**

Chia tập train test theo tỷ lệ 70:30, mà trước đó scale lại ‘Amount’



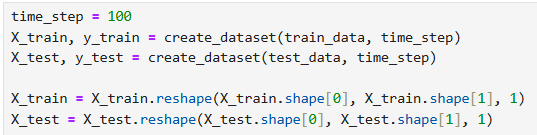
**Hình 6.10: Scale tỉ lệ và chia tập train, test theo tỉ lệ 7:3**

Hàm này chuẩn bị dữ liệu cho mô hình LSTM. Nó tạo ra các chuỗi dữ liệu với độ dài time\_step làm đầu vào (dataX) và giá trị tiếp theo trong chuỗi làm đầu ra (dataY)



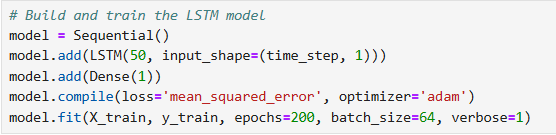
**Hình 6.11: chuẩn bị dữ liệu cho mô hình LSTM**

Tạo dữ liệu đầu vào và đầu ra cho LSTM từ tập train với time\_step là 100



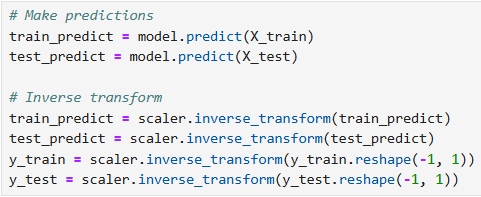
**Hình 6.12: tạo dữ liệu đầu vào và đầu ra cho mô hình LSTM**

Tạo một mô hình tuần tự, huấn luyện mô hình với dữ liệu huấn luyện, 200 epochs, và batch size là 64



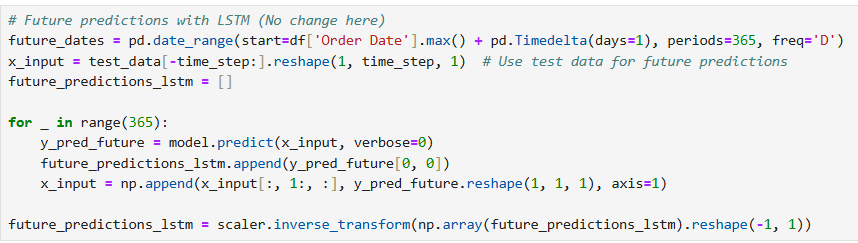
**Hình 6.13: Tạo mô hình tuần tự và huấn luyện mô hình**

Chuyển đổi ngược kết quả dự đoán và dữ liệu thực tế về thang đo ban đầu



**Hình 6.14: Chuyển đổi ngược kết quả dự đoán về thang đo ban đầu**

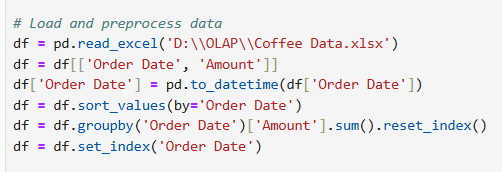
Dự đoán 1 năm sau bằng các tạo các ngày tính từ thời điểm gần nhất của data, thực hiện dự đoán ‘Amount’ theo ngày



**Hình 6.15: Dự đoán doanh thu 1 năm sau từ thời điểm gần nhất**

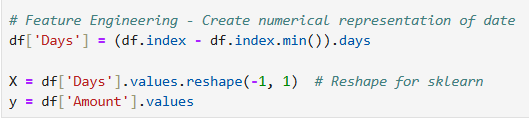
## 2.4. Linear Regression

Tiến hành Tiến hành nhập dữ liệu file csv, đổi định dạng cột ‘Order Date’ thành timestamp, sau đó sắp xếp DataFrame theo cột ‘Order Date’. Nhóm ‘Amount’ theo ngày.



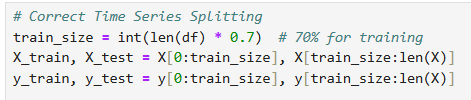
**Hình 6.16: Import và nhóm cột Amount theo ngày**

Tạo cột 'Days' biểu diễn số ngày kể từ ngày đầu tiên trong dữ liệu. Nhằm chuyển đổi dữ liệu thời gian thành dạng số mà mô hình hồi quy tuyến tính có thể sử dụng



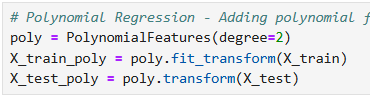
**Hình 6.17: Tạo cột Days để thực hiện mô hình hồi quy tuyến tính**

Chia tập train test theo tỷ lệ 70:30, mà trước đó scale lại ‘Amount’



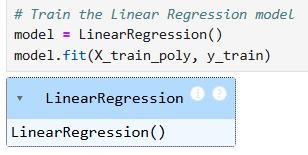
**Hình 6.18: Chia tập train, test theo tỉ lệ 7:3**

Tạo các đặc trưng đa thức bậc 2 từ cột 'Days'. degree=2 nghĩa là tạo các đặc trưng bậc 2 (x, x^2)



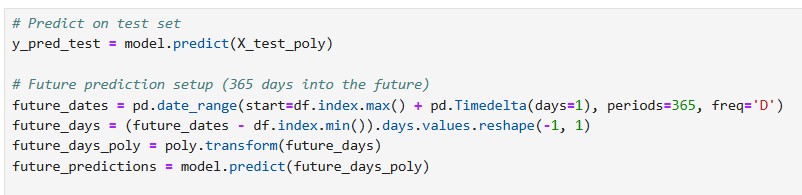
**Hình 6.19: Tạo các đặc trưng đa thức bậc 2**

Tạo mô hình và huấn luyện mô hình



**Hình 6.20: Tạo và huấn luyện mô hình Linear Regression**

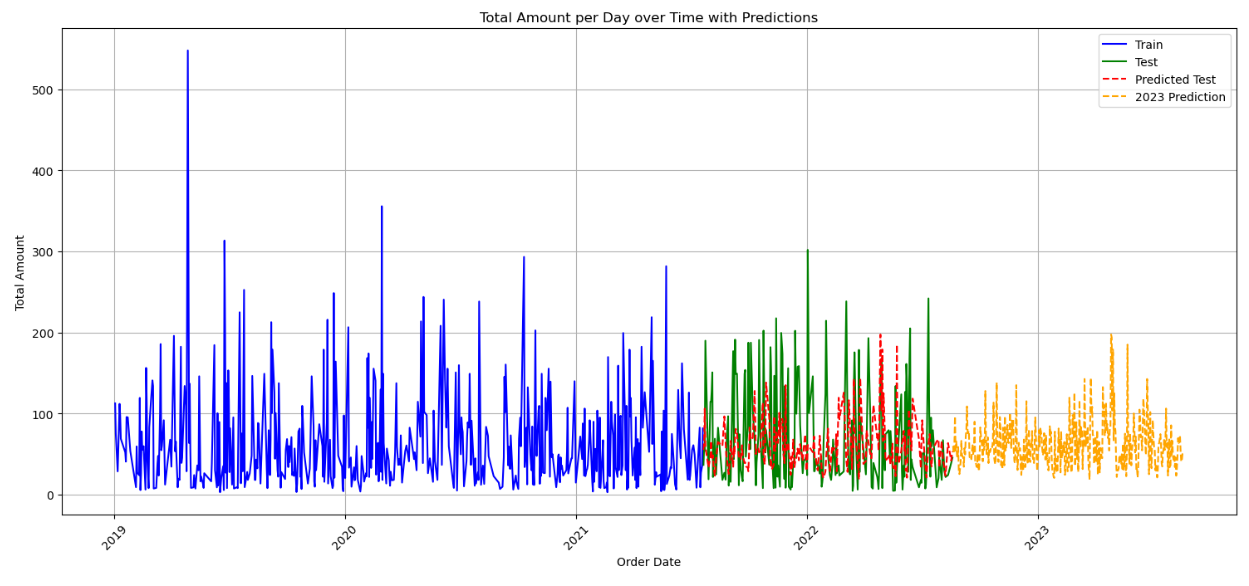
Sau khi dự đoán trên tập test, dự đoán doanh số theo số ngày tương ứng với dữ liệu tương lai được chuyển đổi thành dạng đa thức



**Hình 6.21: Chuyển đổi dữ liệu tương lại thành dạng đa thức**

## 3. Trực quan hóa

### 3.1. Random Forest

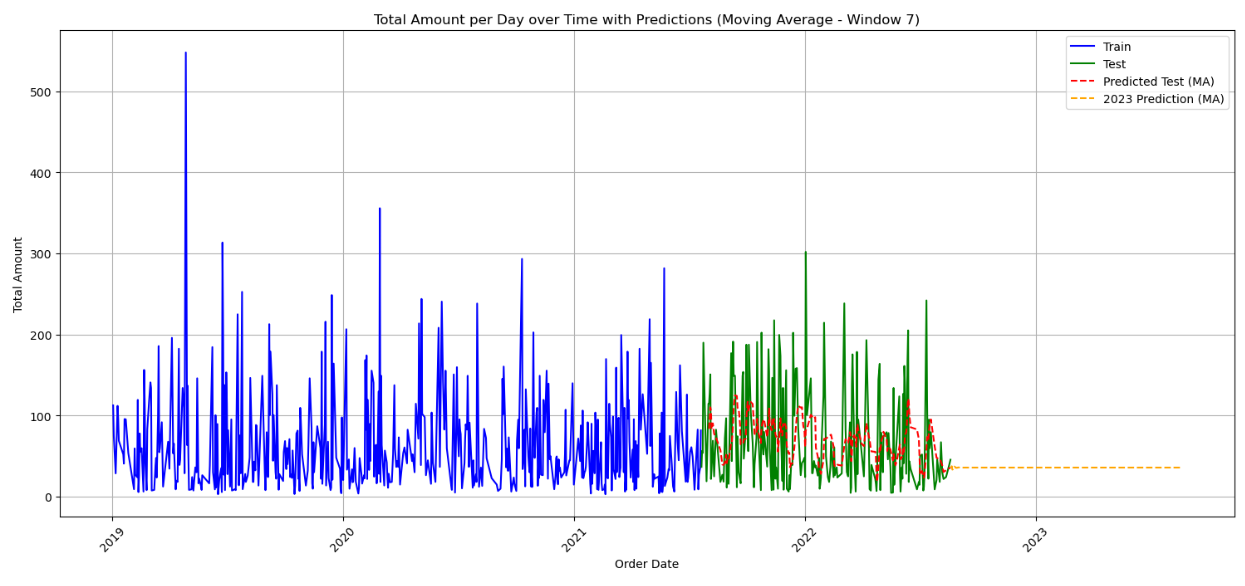


**Hình 6.22: Kết quả dự đoán mô hình Random Forest**

Thông thông số độ sai lệch của thuật toán **Random Forest** là:

* MSE: 4682.95
* RMSE: 68.43
* MAE: 51.90
* MAPE: 161.90

### 3.2. Moving Averages

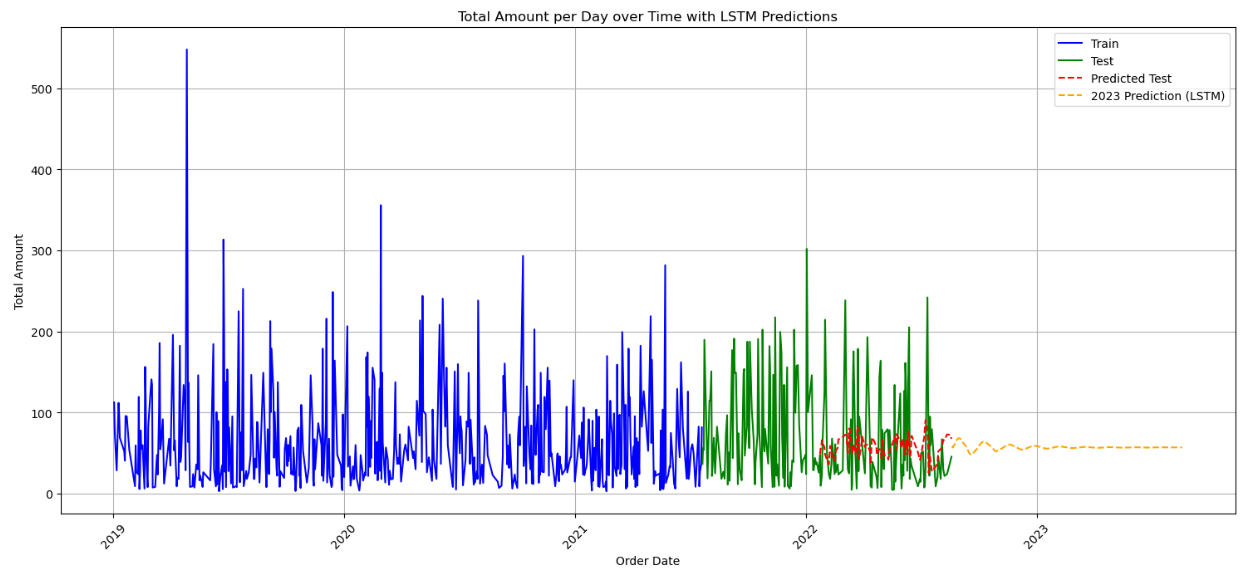


**Hình 6.23: Kết quả dự đoán mô hình Moving Average**

Thông thông số độ sai lệch của thuật toán **Moving Averages** là:

* MSE: 3294.30
* RMSE: 57.39
* MAE: 44.68
* MAPE: 159.98

### 3.3. LSTM

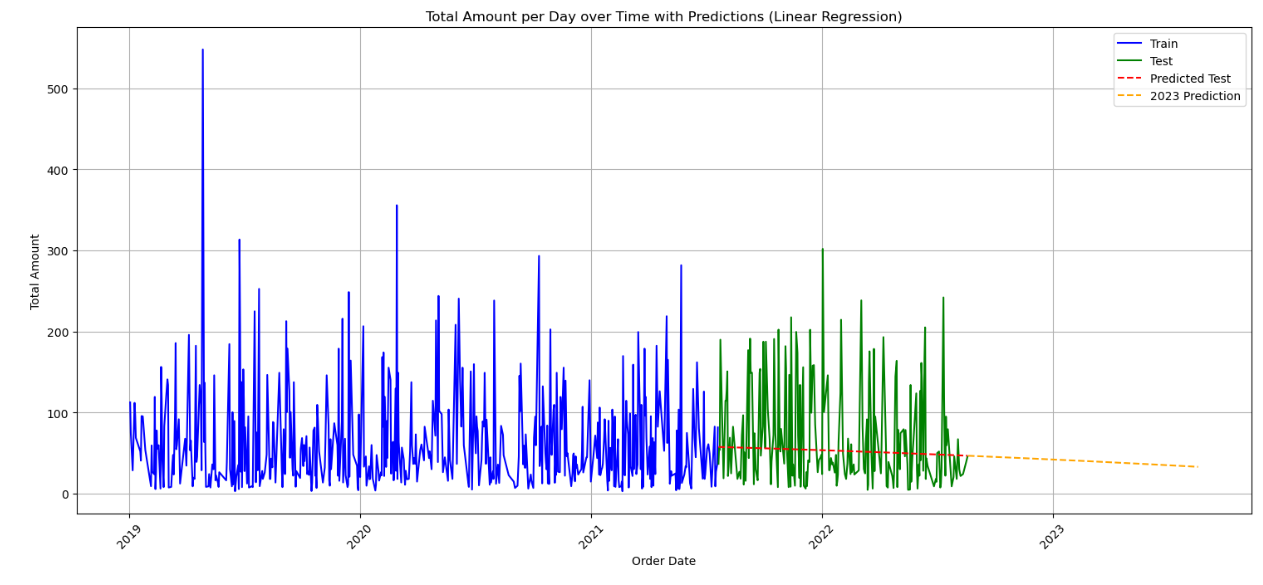


**Hình 6.24: Kết quả dự đoán mô hình LSTM**

Thông thông số độ sai lệch của thuật toán **LSTM** là:

* MSE: 3438.03
* RMSE: 58.63
* MAE: 44.88
* MAPE: 190.10

### 3.4. Linear Regression



**Hình 6.25: Kết quả dự đoán mô hình Linear Regression**

Thông thông số độ sai lệch của thuật toán **Linear Regression** là:

* MSE: 3943.91
* RMSE: 62.80
* MAE: 45.33
* MAPE: 132.22

## 4. Kết quả thực hiện

**Bảng 6.1: Kết quả thông số sai lệch theo từng mô hình**

|  | MSE | RMSE | MAE | MAPE |
| --- | --- | --- | --- | --- |
| Random Forest | 4682.95 | 68.43 | 51.90 | 161.90 |
| **Moving Averages** | **3294.30** | **57.39** | **44.68** | **159.98** |
| LSTM | 3438.03 | 58.63 | 44.88 | 190.10 |
| Linear Regression | 3943.91 | 62.80 | 45.33 | 132.22 |

Theo bảng trên, Moving Averages cho hiệu suất tốt nhất tổng thể (dựa trên MSE, RMSE, MAE): Mô hình Moving Averages đạt được MSE, RMSE và MAE thấp nhất và trong bốn mô hình. Điều này cho thấy rằng, đối với tập dữ liệu cụ thể này, một mô hình đơn giản như Moving Averages có thể hiệu quả hơn các mô hình phức tạp hơn.

# Chương 7. KẾT LUẬN

## 1. Kết quả đạt được

Trong kỳ học vừa qua , nhóm đã tìm hiểu và vận dụng kiến thức về xây dựng kho dữ liệu và OLAP và đạt được các kết quả như sau :

* Nắm rõ các khái niệm cơ bản về kho dữ liệu và OLAP, các tính chất của một kho dữ liệu cần có.
* Năm vững kiến thức và có thể vận dụng, xây dựng một kho dữ liệu hoàn chỉnh dùng để khai thác dữ liệu.
* Trang bị kiến thức về các công cụ SSIS, SSAS, SSRS, Data Mining.
* Xây dựng được kho dữ liệu hoàn chỉnh.
* Hiểu và sử dụng ngôn ngữ truy vấn nhiều chiều MDX.
* Trình bày tối ưu hóa câu truy vấn.

## 2. Hạn chế

Do thời gian hạn ngắn cộng với khối lượng công việc nhiều nên trong quá trình thực hiện đồ án nhóm còn gặp phải một số vấn đề :

* Chưa áp dụng được so sánh kết quả của nhiều phương pháp Data Mining.
* Quá trình SSIS còn rườm rà chưa được tối ưu.
* Các câu truy vấn MDX chưa nâng cao

## 3. Phân công công việc

| **Công việc** | **Hoàng Ngô Thảo Nguyên (100%)** | **Nguyễn Anh Dĩ (100%)** |
| --- | --- | --- |
| Chọn dataset | x | x |
| Tích hợp dữ liệu vào kho (SSIS) | x |  |
| Xây dựng mô hình, khối – Cube (SSAS) | x | x |
| Thực hiện 10 câu truy vấn bằng SSAS, Pivot table, MDX | x |  |
| Data mining |  | x |
| Report | x | x |
| Viết báo cáo | x | x |
| Thiết kế slide |  | x |
| Quay Video | x | x |