BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**

****

**ĐỒ ÁN KẾT THÚC MÔN HỌC**

**MÔN: LẬP TRÌNH R CHO PHÂN TÍCH**

**ĐỀ TÀI**

**PHÂN TÍCH VÀ XÂY DỰNG MÔ HÌNH DỰ ĐOÁN VỚI TẬP DỮ LIỆU BÁN HÀNG PIZZA RESTAURANT SALES**

**GVHD: TS Phan Thị Thể**

**Mã học phần:** 241RPAN233577\_01

|  |  |
| --- | --- |
| Nguyễn Thị Ngọc Hân | MSSV: 22133017 |
| Võ Triệu Phúc | MSSV: 22133043 |
| Nguyễn Thị Hồng Thơ | MSSV: 22151305 |
| Nguyễn Hoàng | MSSV: 22133020 |

**Sinh viên thực hiện:**

*Tp Hồ Chí Minh, tháng 11 năm 2024*

|  |  |
| --- | --- |
| BỘ GIÁO DỤC VÀ ĐÀO TẠO  **TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT**  **THÀNH PHỐ HỒ CHÍ MINH** | **CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**  **Độc lập – Tự do – Hạnh phúc** |

**NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN**

**Họ và tên sinh viên thực hiện:**

* Nguyễn Thị Ngọc Hân - 22133020
* Võ Triệu Phúc – 22133043
* Nguyễn Thị Hồng Thơ - 22151305
* Nguyễn Hoàng - 22133020

**Chuyên ngành:** Kỹ thuật dữ liệu (Data Engineering)

**Đề tài:** Phân tích và xây dựng mô hình dự đoán với tập dữ liệu bán hàng pizza restaurant sales

**Môn học:** Lập Trình R Cho Phân Tích

**Nhận xét:**

................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

................................................................................................................................................................................................................................................................................................

Tp HCM, / / 2024

Giảng viên hướng dẫn

(Họ tên và chữ ký)

**MỤC LỤC**

[**PHẦN 1. MỞ ĐẦU 6**](#_Toc183980331)

[**1.1. Tóm tắt 6**](#_Toc183980332)

[**1.2. Giới thiệu 6**](#_Toc183980333)

[**1.3. Mục tiêu nghiên cứu 7**](#_Toc183980334)

[**PHẦN 2. NỘI DUNG 8**](#_Toc183980335)

[**2.1. Thông tin về tập dữ liệu 8**](#_Toc183980336)

[**2.2. Tiền xử lý dữ liệu 9**](#_Toc183980337)

[**2.3. Phân tích và trực quan hóa dữ liệu 13**](#_Toc183980338)

[**2.3.1. Phân tích tổng quan 13**](#_Toc183980339)

[**2.3.2. Phân tích nhu cầu tiêu thụ pizza theo thời gian 15**](#_Toc183980340)

[**2.3.3. Phân tích nguyên liệu 24**](#_Toc183980341)

[**2.3.3.1. Phân tích tần suất xuất hiện các từ trong nguyên liệu làm pizza 24**](#_Toc183980342)

[**2.3.3.2. Phân tích top 10 nguyên liệu sử dụng nhiều nhất theo tháng 25**](#_Toc183980343)

[**2.3.4. Phân tích doanh số 27**](#_Toc183980344)

[**2.3.4.1. Phân tích doanh số theo kích cỡ bánh 27**](#_Toc183980345)

[**2.3.4.2. Phân tích doanh số theo danh mục bánh 28**](#_Toc183980346)

[**2.3.4.3. Phân tích doanh số bán hàng của từng bánh pizza 30**](#_Toc183980347)

[**2.3.5. Phân tích doanh thu theo thời gian 35**](#_Toc183980348)

[**2.3.5.1. Phân tích doanh thu theo giờ trong ngày của một tuần 35**](#_Toc183980349)

[**2.3.5.2. Phân tích doanh thu một ngày dùng Time Series – Chuỗi thời gian 37**](#_Toc183980350)

[**2.3.5.3. Phân tích doanh thu theo tháng 42**](#_Toc183980351)

[**2.4. Xây dựng mô hình dự đoán 44**](#_Toc183980352)

[**2.4.1. Dự đoán doanh số theo sản phẩm 44**](#_Toc183980353)

[**2.4.1.1. Dự đoán doanh số của sản phẩm bằng hồi quy tuyến tính 46**](#_Toc183980354)

[**2.4.1.2. Dự đoán doanh số theo sản phẩm bằng Random Forest 47**](#_Toc183980355)

[**2.4.1.3. Dự đoán doanh số bằng mô hình GBM 53**](#_Toc183980356)

[**2.4.1.4. Kết luận 61**](#_Toc183980357)

[**2.4.2. Dự đoán doanh thu trong 6 tháng tiếp theo bằng mô hình ARIMA 61**](#_Toc183980358)

[**PHẦN 3. KẾT LUẬN 80**](#_Toc183980359)

[**PHẦN 4. TÀI LIỆU THAM KHẢO 81**](#_Toc183980360)

[**PHẦN 5. BẢNG PHÂN CÔNG VÀ ĐÁNH GIÁ 83**](#_Toc183980361)

[**5.1. Bảng phân công nhiệm vụ 83**](#_Toc183980362)

[**5.2. Thành viên nhận xét và đánh giá 83**](#_Toc183980363)

# MỞ ĐẦU

## Tóm tắt

Nghiên cứu này dùng ngôn ngữ R để phân tích và xây dựng mô hình dự đoán trên tập dữ liệu bán hàng pizza trong quá khứ. Các biểu đồ Histogram, Boxplot, Line chart, Bar chart, Pie chart, Scatter plot và WordCloud được ứng dụng để phân tích dữ liệu và trực quan hóa xu hướng trong nhu cầu pizza theo thời gian, tìm ra các thành phần nguyên liệu được thường xuyên sử dụng, thể hiện doanh số, doanh thu của cửa hàng… Đối với doanh số, kiểm định Kruskal-Wallis được dùng để kiểm tra sự biến động doanh số của từng pizza giữa các tháng có đáng kể không và ANOVA được dùng để so sánh sự chênh lệch doanh số giữa các danh mục bánh. Sau đó các mô hình khác nhau đã được sử dụng để thực hiện dự đoán doanh số: Mô hình Hồi quy tuyến tính được chọn do tính đơn giản và khả năng giải thích dễ dàng, trong khi Rừng ngẫu nhiên và GBM được chọn để xử lý các mối quan hệ phức tạp giữa các biến độc lập. Bên cạnh đó, mô hình ARIMA được dùng để dự đoán doanh thu cho 6 tháng tiếp theo.

## Giới thiệu

Dữ liệu được ví như “dầu mỏ của thế kỷ 21” nhờ khả năng thúc đẩy sự đổi mới và cải tiến trong nhiều lĩnh vực. Giá trị của dữ liệu không chỉ nằm ở bản thân các con số, mà còn ở những thông tin có thể khai thác và ứng dụng từ dữ liệu. Đặc biệt trong ngành F&B, khi được thu thập, phân tích và diễn giải đúng cách, dữ liệu bán hàng và hành vi của người tiêu dùng trở thành nguồn tài nguyên chiến lược quan trọng giúp doanh nghiệp đưa ra những quyết định chính xác, giảm thiểu rủi ro và thúc đẩy tăng trưởng. Hơn nữa, phân tích dữ liệu bán hàng sẽ giúp tối ưu hóa hàng tồn kho của doanh nghiệp, giảm tình trạng dư thừa, cũng như có thể tăng số lượng các mặt hàng phổ biến để tránh tình trạng thiếu hàng (sold out) và tối đa hóa doanh số cho doanh nghiệp.

Cùng với những mục đích trên, trong nghiên cứu này, nhóm chúng em thực hiện phân tích tập dữ liệu bán hàng năm 2015 của một cửa hàng pizza và phát triển các mô hình dự đoán doanh số và doanh thu. Từ đó, có thể rút ra những nhận định hữu ích để hỗ trợ việc ra quyết định kinh doanh, tăng cường hiệu quả hoạt động và xây dựng các chiến lược marketing phù hợp.

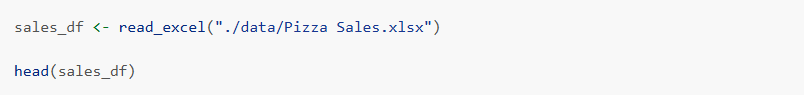
## Mục tiêu nghiên cứu

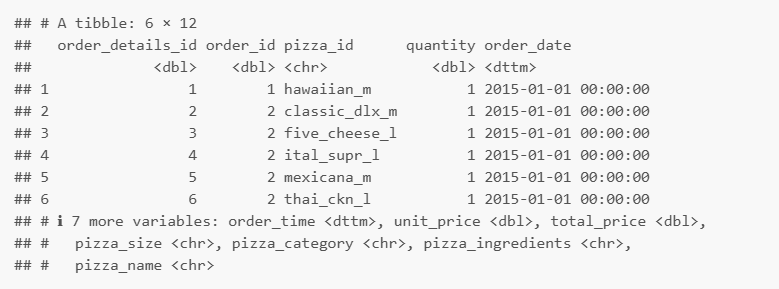
Để đạt được mục đích đã nêu trên, nhóm chúng em đặt ra các mục tiêu sau:

* Phân tích và trực quan hóa nhu cầu tiêu thụ pizza theo thời gian, xác định rõ xu hướng tiêu thụ theo các giai đoạn cụ thể như mùa, tháng, tuần, ngày trong tuần và giờ trong ngày.
* Phân tích và trực quan hóa tần suất sử dụng, xác định các nguyên liệu dùng nhiều nhất mỗi tháng.
* Phân tích và trực quan hóa doanh số theo kích cỡ bánh, danh mục bánh, xác định các pizza bán chạy nhất cũng như kém phổ biến nhất trong năm 2015. Xác định các yếu tố ảnh hưởng đến doanh số.
* Xây dựng mô hình dự đoán doanh số của sản phẩm có độ chính xác trên 80%.
* Phân tích và trực quan hóa doanh thu theo thời gian: doanh thu theo giờ, doanh thu theo ngày, doanh thu theo tháng của cửa hàng trong năm 2015.
* Xây dựng mô hình dự đoán doanh thu trong 6 tháng tiếp theo của cửa hàng.

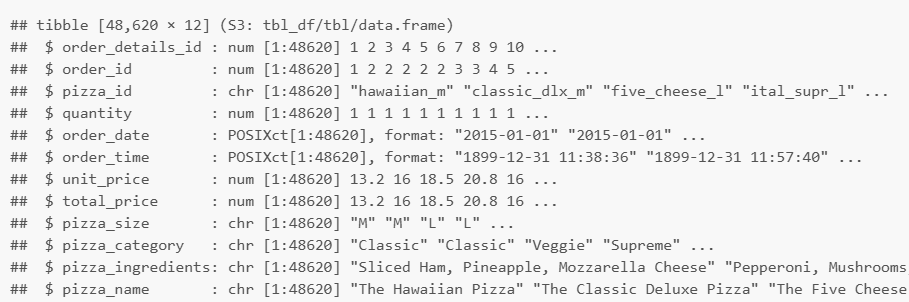
# NỘI DUNG

## Thông tin về tập dữ liệu









Tập dữ liệu Pizza Sales có tổng cộng 48,620 dòng và 12 cột, cung cấp thông tin chi tiết về các đơn hàng bán pizza. Mỗi dòng tương ứng với một mục trong đơn hàng, được xác định bởi order\_details\_id, và thuộc về một đơn hàng nhận diện qua order\_id. Dữ liệu bao gồm các thông tin quan trọng như ngày và giờ đặt hàng, loại pizza, kích cỡ, số lượng, đơn giá… Cụ thể về thông tin của 12 biến và kiểu dữ liệu tương ứng như sau:

* order\_details\_id: Kiểu số (numeric). Đây là mã định danh duy nhất cho từng dòng trong tập dữ liệu, đại diện cho mỗi chi tiết đơn hàng.
* order\_id: Kiểu số. Đây là mã định danh của mỗi đơn hàng. Nhiều dòng có order\_id giống nhau khi chúng cùng thuộc về một đơn hàng.
* pizza\_id: Kiểu chuỗi ký tự (character). Đây là mã nhận diện từng pizza.
* quantity: Kiểu số, cho biết số lượng pizza được đặt trong một chi tiết đơn hàng.
* order\_date: có kiểu POSIXct, dữ liệu được lưu dưới dạng số nguyên biểu thị số giây kể từ epoch time (01/01/1970), và được hiển thị theo dưới dạng yyyy-mm-dd, cho biết ngày đặt hàng.
* order\_time: tương tự như order\_date, biến này có kiểu POSIXct, cho biết thời gian đặt hàng. Nhưng dữ liệu có dạng YYYY-mm-dd hh:mm:ss, trong đó ngày tháng năm là không có nghĩa nên cần xử lý trước khi phân tích.
* unit\_price: Kiểu số, cho biết đơn giá của một chiếc bánh pizza cụ thể.
* total\_price: Kiểu số, cho biết tổng số tiền của dòng chi tiết đơn hàng đó, bằng unit\_price × quantity.
* pizza\_size: Kiểu chuỗi ký tự, cho biết kích cỡ pizza.
* pizza\_category: Kiểu chuỗi ký tự, cho biết pizza được mua thuộc danh mục nào.
* pizza\_ingredients: Kiểu chuỗi ký tự, cho biết các nguyên liệu làm nên pizza đó.
* pizza\_name: Kiểu chuỗi ký tự, cho biết tên của pizza.

## Tiền xử lý dữ liệu

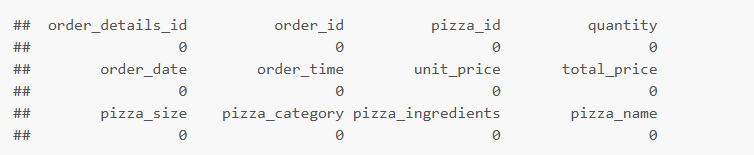
Sao chép dữ liệu vào biến df để xử lý trên bản sao không gây ảnh hưởng đến dữ liệu gốc.



**Kiểm tra giá trị thiếu (na) và trùng lặp (duplicate)**

* Kiểm tra xem có tồn tại giá trị thiếu (na) ở các cột hay không





Không tồn tại giá trị na.

* Kiểm tra xem có tồn tại duplicate hay không





Không có quan sát nào bị trùng lặp

**Chuyển đổi cột pizza\_size, pizza\_category thành factor**

Xem các giá trị duy nhất của cột pizza\_size và pizza\_category để xác định thành phần trong levels của factor

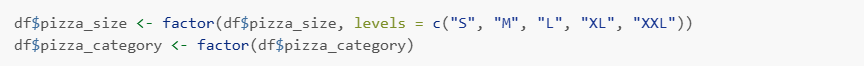






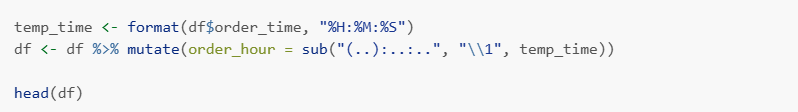


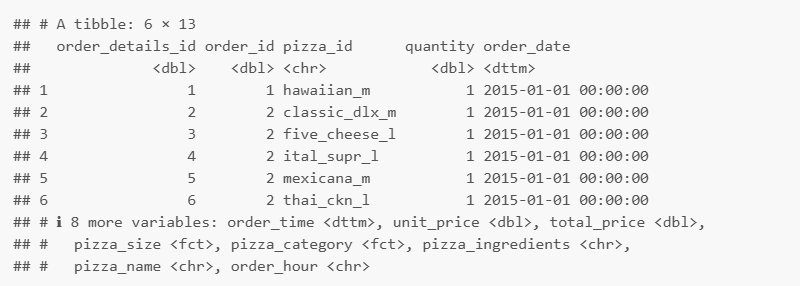
Thực hiện chuyển đổi thành factor



**Tạo cột order\_hour**

Dữ liệu của cột order\_time là kiểu POSIXct, được lưu trữ dưới dạng số thực đơn vị giây tính từ ngày 1 tháng 1 năm 1970, 00:00:00 UTC, và hiển thị dưới dạng yyyy-mm-dd HH:mm:ss. Nhằm phục vụ mục đích phân tích, em sẽ dựa vào cột này để tạo cột order\_hour thể hiện giờ đặt bánh của khách.



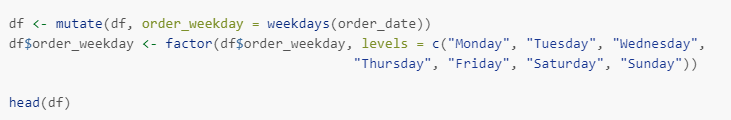


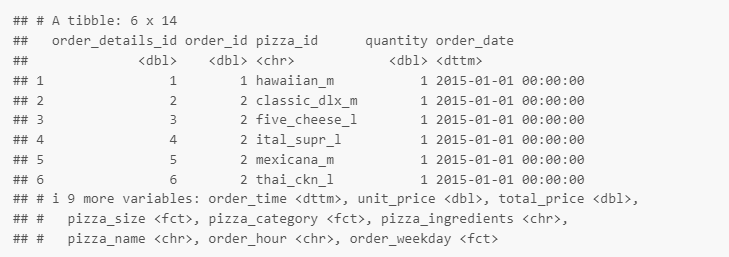
**Tạo cột order\_weekday**

Thêm cột mới tên order\_weekday có kiểu factor cho biết khách hàng đã đặt đơn vào thứ mấy.





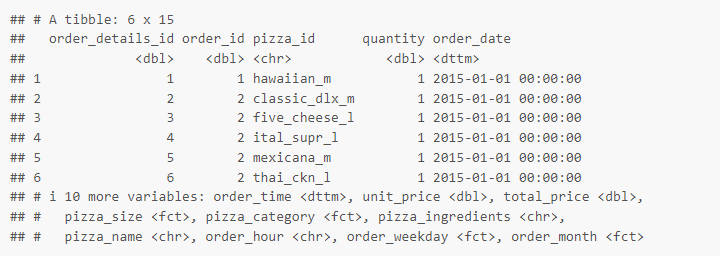




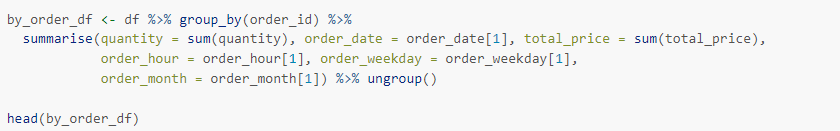
**Tạo cột order\_month**

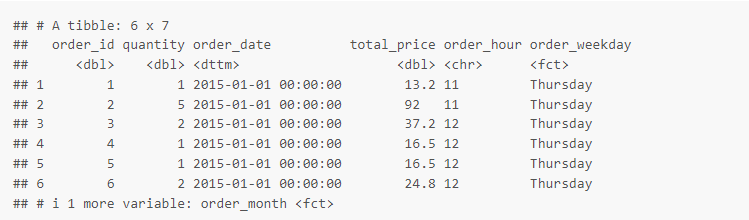
Thêm cột mới tên order\_month có kiểu factor cho biết khách hàng đã đặt đơn vào tháng nào.





**Tạo data frame chứa dữ liệu của từng đơn đặt hàng**

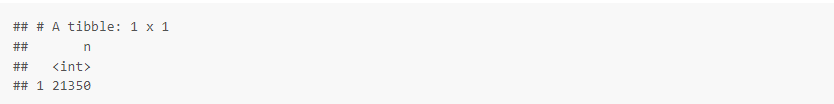




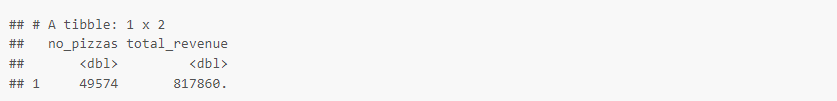
## Phân tích và trực quan hóa dữ liệu

### Phân tích tổng quan







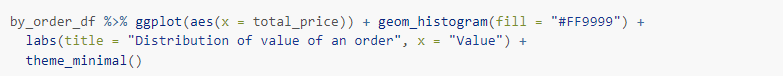


Trong năm 2015, cửa hàng có 21350 đơn đặt hàng, bán ra tổng 49574 chiếc bánh pizza với tổng doanh thu là 817860 USD.

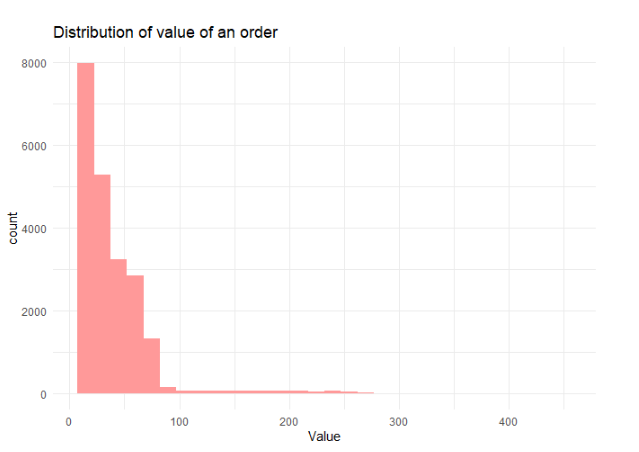
Trung bình, mỗi đơn hàng đặt khoảng 2 chiếc bánh pizza, thể hiện phần lớn khách hàng có xu hướng đặt cho gia đình nhỏ hoặc nhóm bạn. Số lượng pizza lớn nhất cho một đơn hàng là 28 chiếc, cho thấy có thể có những đơn đặt hàng số lượng lớn như từ các sự kiện, các buổi tiệc.

Dữ liệu cũng cho thấy sự phân hóa trong chi tiêu của khách hàng: đơn hàng thấp nhất là 9.75 USD, cao nhất là 444.4 USD, trong khi trung bình là 32.5 USD.

*Cửa hàng có thể tận dụng sự đa dạng trong nhóm khách hàng để tạo các chiến lược khuyến mãi phù hợp cho từng nhóm và thiết kế thực đơn, combo hợp lý.*



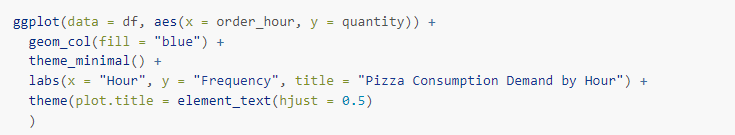


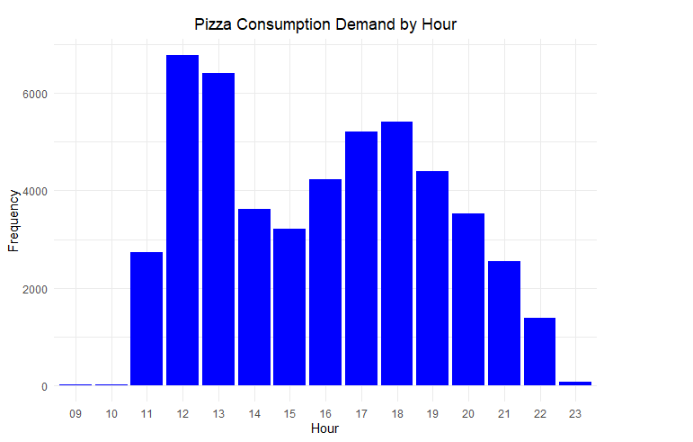


*Biểu đồ histogram của giá trị đơn hàng cho thấy phân phối giá trị lệch phải. Số lượng đơn hàng giá trị hơn 100 USD rất ít. Phần lớn đơn hàng trị giá từ 10 đến 20 USD. Xu hướng này cho thấy cửa hàng chủ yếu thu hút khách hàng với các đơn hàng có giá trị nhỏ và trung bình, nhưng vẫn có tiềm năng phục vụ các đơn hàng có giá trị lớn.*

### Phân tích nhu cầu tiêu thụ pizza theo thời gian

**Biểu đồ nhu cầu tiêu thụ pizza theo giờ:**

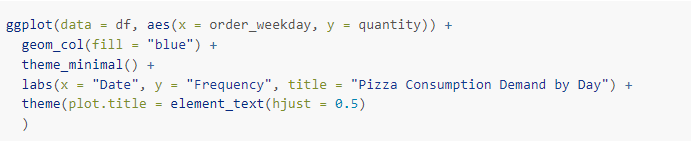


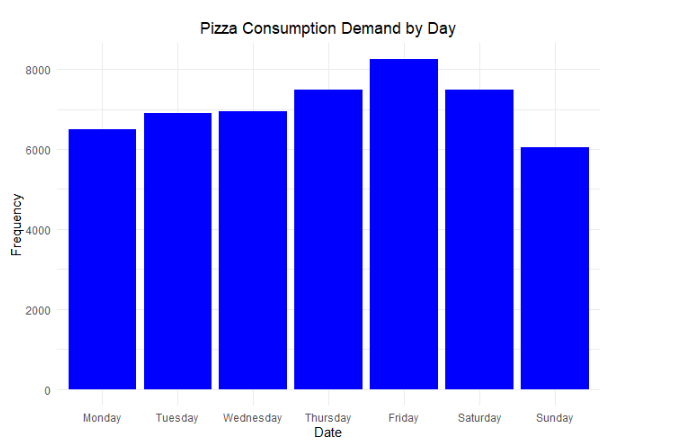


* Biểu đồ cho thấy sự phân bố số lượng đơn đặt hàng theo từng giờ trong ngày.
* Giờ có lượng đơn hàng cao nhất là từ 12h đến 13h trưa. Thời gian từ 17h đến 18h cũng có lượng đơn hàng khá cao. Hai khoảng giờ cao điểm này có thể là các thời điểm mà người tiêu dùng có nhiều thời gian rảnh sau giờ làm việc cho bữa ăn trưa và ăn tối. Các khung giờ khác có số lượng đơn hàng giảm dần và thấp nhất là vào khoảng 9h,10h sáng và 23h khuya.

*Dựa vào biểu đồ này, ta có thể quan sát rõ hành vi của người tiêu dùng. Giờ trưa và giờ tan làm có thể là lúc người tiêu dùng có thời gian và nhu cầu đặt thức ăn nhiều hơn. Điều này giúp nhà hàng lập kế hoạch hỗ trợ và chiến lược quảng cáo tốt hơn để thu hút và giữ chân khách hàng.*

**Biểu đồ nhu cầu tiêu thụ pizza theo ngày:**

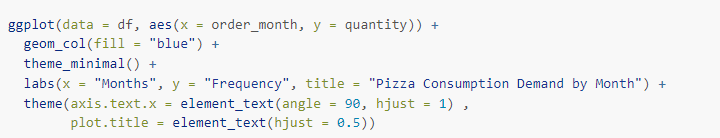


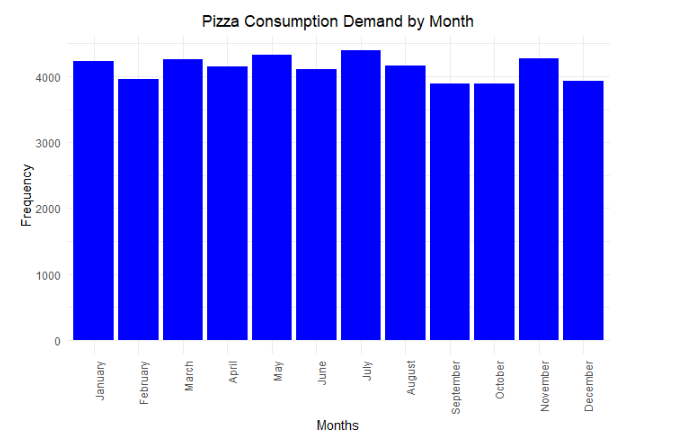


Thứ Sáu có nhu cầu cao nhất. Điều này cho thấy mọi người có xu hướng đặt pizza nhiều hơn khi cuối tuần đến gần. Còn lại các ngày khác trong tuần, từ Thứ Hai đến Thứ Năm, nhu cầu tương đối ổn định cho thấy sự tiêu thụ pizza đều đặn trong suốt các ngày làm việc.Nhu cầu thấp nhất vào Chủ Nhật, có khả năng cao khách hàng chuẩn bị cho tuần làm việc sắp tới.

*Biểu đồ cho thấy nhu cầu tiêu thụ pizza tăng cao vào cuối tuần, đạt đỉnh vào Thứ Sáu, có thể do các buổi tụ họp xã hội hoặc xu hướng ăn uống ngoài khi bắt đầu cuối tuần.*

**Biểu đồ nhu cầu tiêu thụ pizza theo tháng:**



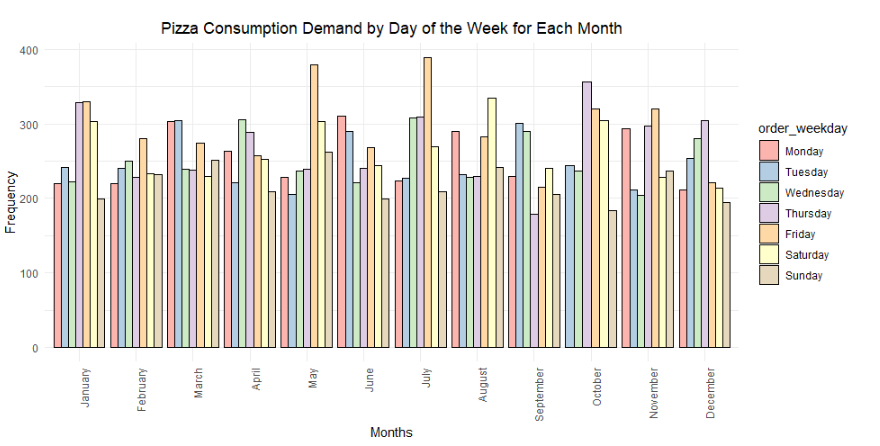


* Các tháng có nhu cầu cao nhất là Tháng Bảy và Tháng Mười Một, cho thấy những khoảng thời gian này có thể có các sự kiện đặc biệt hoặc kỳ nghỉ lễ khiến nhu cầu tăng lên.
* Tháng Hai và Tháng Chín có nhu cầu tiêu thụ pizza thấp hơn so với các tháng khác, có thể do đây là khoảng thời gian sau các kỳ nghỉ lễ hoặc đầu mùa học tập và làm việc.

*Nhìn chung, mức tiêu thụ pizza trong các tháng khá đồng đều, chỉ có sự chênh lệch nhẹ giữa các tháng. Điều này cho thấy pizza của cửa hàng thu hút thực khách quanh năm. Nhu cầu tiêu thụ pizza dường như ổn định trong suốt cả năm, với một số tháng có nhu cầu cao hơn, đặc biệt là vào mùa hè (Tháng Bảy) và vào mùa lễ (Tháng Mười Một). Sự tăng giảm nhẹ giữa các tháng có thể phản ánh sự thay đổi trong thói quen ăn uống của người tiêu dùng theo mùa và các dịp lễ.*

**Biểu đồ nhu cầu tiêu thụ pizza theo ngày trong tuần của mỗi tháng:**

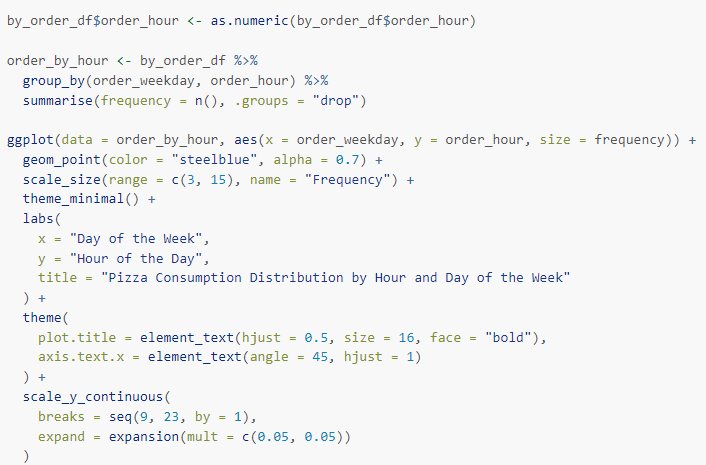


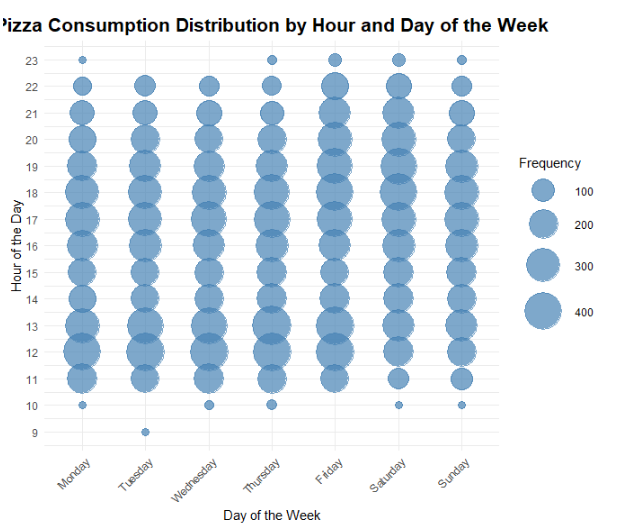


* Sự phổ biến vào cuối tuần: Trong hầu hết các tháng, Thứ Sáu và Thứ Bảy có xu hướng có tần suất đặt hàng cao hơn, cho thấy nhu cầu tiêu thụ pizza tăng vào cuối tuần, có thể do đây là thời điểm mọi người có thời gian rảnh rỗi hoặc tổ chức các buổi tụ họp.
* Biến động theo tháng: Tháng Năm, Tháng Bảy, và Tháng Mười có một số ngày có tần suất cao hơn rõ rệt (đặc biệt vào các ngày cuối tuần), có thể do những tháng này có các sự kiện, kỳ nghỉ hoặc thời tiết thuận lợi cho các hoạt động ăn uống ngoài trời.
* Nhu cầu ổn định giữa tuần: Nhu cầu tiêu thụ pizza vào các ngày giữa tuần (Thứ Hai đến Thứ Năm) ổn định hơn, với ít biến động giữa các tháng. Điều này cho thấy pizza vẫn được tiêu thụ đều đặn nhưng không quá cao vào những ngày đầu tuần và giữa tuần.
* Chênh lệch theo ngày: Sự khác biệt về tần suất giữa các ngày trong tuần lớn nhất vào những tháng mùa hè và mùa thu, cho thấy nhu cầu có thể bị ảnh hưởng bởi mùa hoặc các kỳ nghỉ kéo dài.

*Ta có thể kết luận rằng nhu cầu tiêu thụ pizza thay đổi theo cả tháng và ngày trong tuần, với xu hướng tiêu thụ cao vào các ngày bắt đầu cuối tuần và một số tháng cụ thể có kì nghỉ lễ hoặc thời tiết đẹp. Điều này có thể được sử dụng để dự đoán nhu cầu và tối ưu hóa nguồn cung cho các khoảng thời gian cao điểm trong năm.*

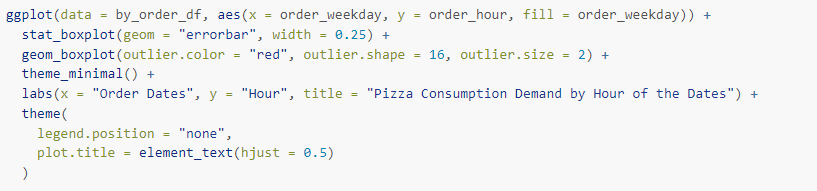
**Biểu đồ nhu cầu tiêu thụ pizza theo giờ và ngày trong tuần:**

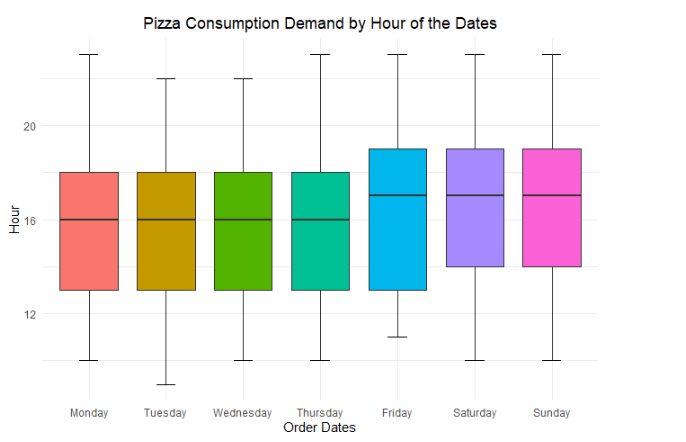




Các đơn đặt hàng chủ yếu được đặt từ trưa (11 giờ) đến tối (22 giờ), biểu hiện khoảng thời gian từ trưa đến tối là thời gian cao điểm cho nhu cầu tiêu thụ pizza, đặc biệt vào Thứ Sáu và Thứ Bảy. Ngoài khung giờ này là các điểm dữ liệu ngoài (outliers) cho các đơn hàng vào sáng sớm hoặc đêm muộn. Điều này cho thấy có một số ít đơn đặt hàng trong khoảng thời gian ngoài giờ cao điểm.

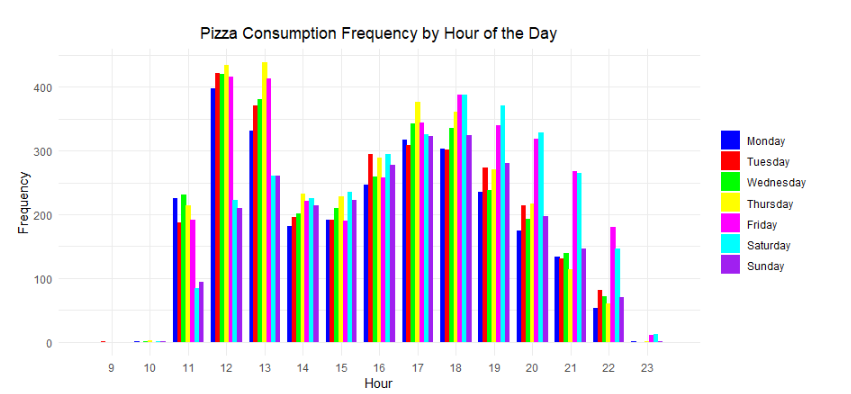
*Có thể thấy ít có sự biến động giữa các ngày trong tuần và xu hướng chung là tiêu thụ pizza ổn định từ buổi trưa đến tối. Không có ngày nào có sự biến động lớn hoặc bất thường. Nhu cầu khá ổn định trong cả tuần nhưng có xu hướng cao hơn vào Thứ Sáu và Thứ Bảy*





Biểu đồ hộp (boxplot) thể hiện sự phân phối thời gian trong ngày mà pizza được tiêu thụ, được chia theo từng ngày trong tuần. Phân phối phổ biến của thời gian đặt hàng là từ 13h đến 19h tối. Thời gian trung bình đặt hàng các ngày trong tuần là 16h và các ngày cuối tuần (Friday, Saturday, Sunday) có trung vị cao hơn là 17h. Điều này có thể phản ánh vào cuối tuần nhu cầu cao hơn vào buổi tối hoặc các khung giờ muộn. Đặc biệt là Friday, có độ rộng của box lớn nhất, hay mức độ phân tán rộng hơn các ngày khác thể hiện rằng khách hàng đặt pizza vào cả những khung giờ không phổ biến (ngoài giờ cao điểm).

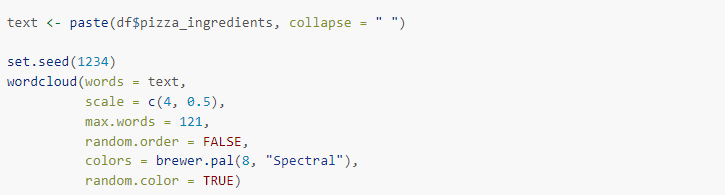




* Thời gian cao điểm: Các giờ cao điểm tiêu thụ pizza tập trung vào buổi trưa (11 giờ - 13 giờ) và buổi tối (17 giờ - 19 giờ). Đỉnh cao nhất của nhu cầu là vào khoảng 12 giờ trưa, với tần suất đặt hàng cao nhất vượt qua 900 đơn vào một số ngày trong tuần, đặc biệt vào các ngày Thứ Ba, Thứ Năm, và Thứ Sáu.
* Sự khác biệt theo ngày trong tuần:
* Thứ Ba và Thứ Sáu: Có tần suất tiêu thụ pizza cao vào giờ trưa (11 giờ - 13 giờ) và buổi tối (17 giờ - 19 giờ).
* Thứ Bảy: Tần suất đặt hàng tăng cao vào buổi tối từ 17 giờ đến 19 giờ, có thể do đây là ngày cuối tuần.
* Chủ Nhật: Tần suất tiêu thụ pizza giảm vào buổi trưa nhưng vẫn có mức độ đặt hàng tương đối ổn định vào buổi tối.
* Thời gian ít đơn đặt hàng: Trong các khoảng thời gian sáng sớm (trước 10 giờ) và đêm muộn (sau 21 giờ), tần suất tiêu thụ pizza khá thấp ở tất cả các ngày. Đặc biệt là từ 22 giờ đến 23 giờ, nhu cầu tiêu thụ pizza gần như không đáng kể.

### Phân tích nguyên liệu

#### Phân tích tần suất xuất hiện các từ trong nguyên liệu làm pizza

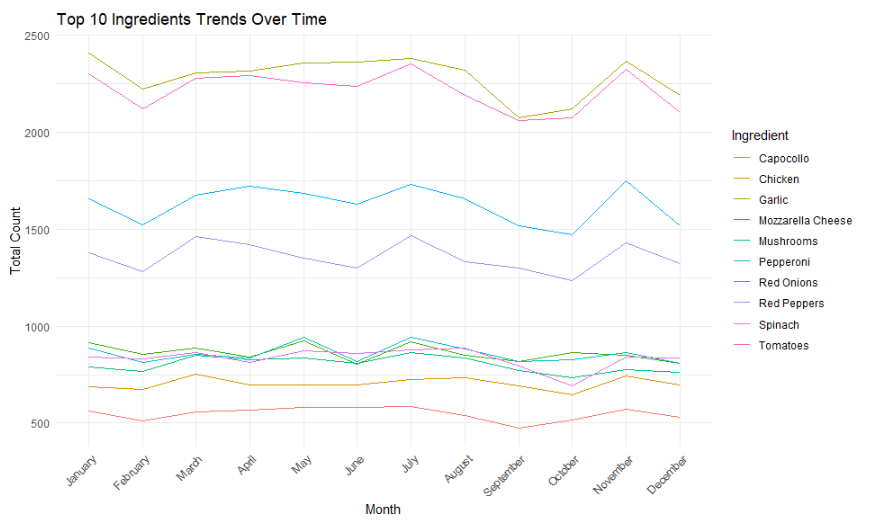




* Từ có tần suất lớn nhất là “cheese”, một thành phần quan trọng và phổ biến nhất trong pizza. Bên cạnh đó còn có các từ như “garlic”, “tomatoes”, “peppers”,..đây cũng là những nguyên liệu phổ biến. Điều này phù hợp với đặc trưng của món pizza.
* Ngoài ra, một số nguyên liệu ít phổ biến với kích thước từ nhỏ và các loại phô mai đặc biệt (gouda, ricotta) có thể phù hợp với những khách hàng thích hương vị độc đáo nên tần suất rất ít.

#### Phân tích top 10 nguyên liệu sử dụng nhiều nhất theo tháng





Garlic là nguyên liệu chủ lực và phổ biến nhất trong công thức pizza, vượt trội hơn các thành phần khác về số lần sử dụng. Đây có thể do hương vị mạnh mẽ và phù hợp với nhiều loại pizza khác nhau. Vị trí thứ hai của Tomatoes phản ánh tầm quan trọng của nguyên liệu này trong nước sốt pizza. Xếp tiếp theo là Red Onion và Red Pepper Mozzarella Cheese cũng có tần suất phổ biến. Các nguyên liệu đứng đầu này có xu hướng biến động thấp nhất vào tháng Hai, tháng Mười, tăng mạnh vào mùa hè (tháng Bảy) và mùa đông (tháng Mười Một).

*Nhà hàng có thể dựa vào phân tích để cân nhắc quản lý kho nguyên liệu phù hợp, đặc biệt là các thành phần phổ biến trong các món đặc trưng. Đồng thời, có thể thử nghiệm các loại pizza mới với các nguyên liệu ít phổ biến hơn để thu hút khách hàng thích trải nghiệm hương vị mới.*

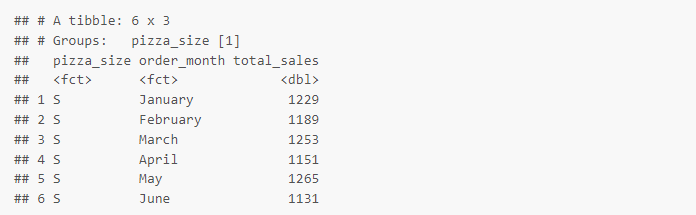
### Phân tích doanh số

#### Phân tích doanh số theo kích cỡ bánh

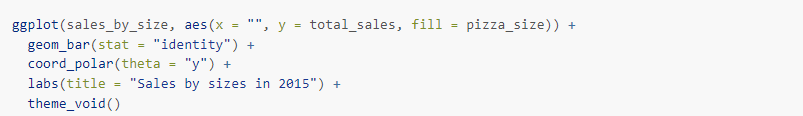
* Tổng hợp dữ liệu tính doanh số mỗi tháng của từng loại pizza

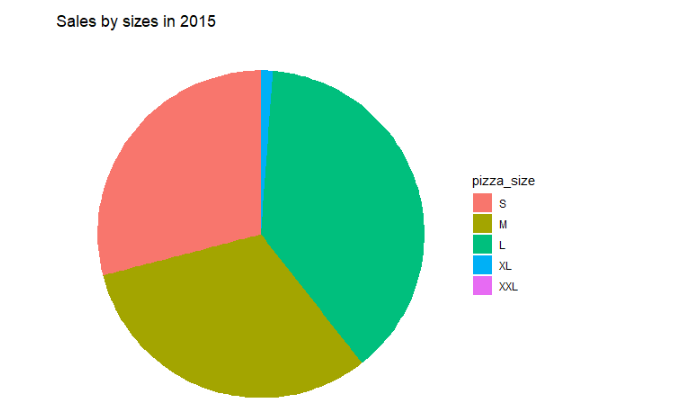






* Trực quan hóa bảng số liệu trên bằng biểu đồ tròn

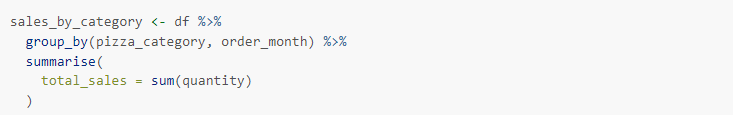




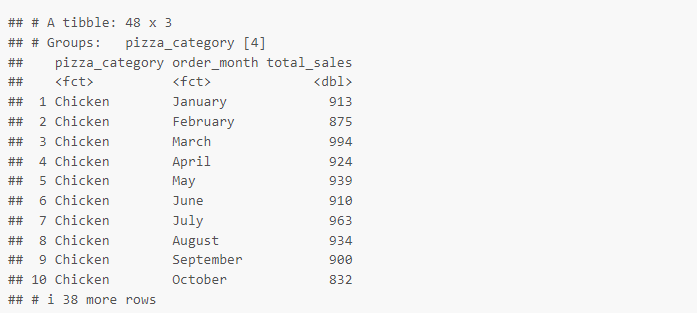
Bánh pizza cỡ L bán chạy nhất với 18956 chiếc bánh chiếm 38% doanh số, theo sau là cỡ M với 15635 bánh  (32%) và cỡ S với 14403 bánh (29%). Bánh size XL và XXL có doanh số rất thấp, cả hai chỉ chiếm chưa tới 2% tổng doanh số của năm 2015.

#### Phân tích doanh số theo danh mục bánh

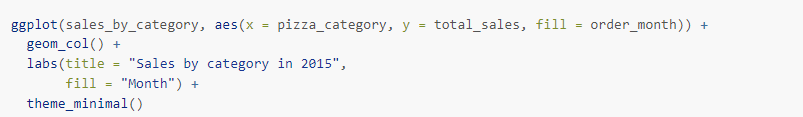
* Tổng hợp dữ liệu theo danh mục bánh

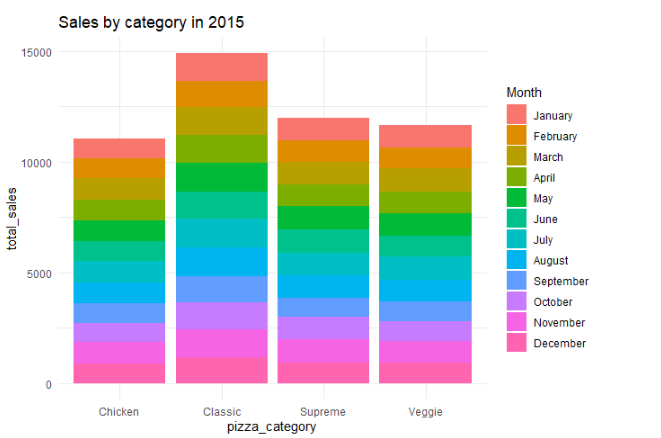






* Trực quan hóa



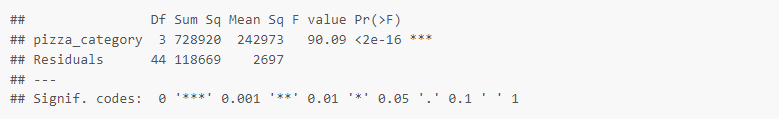


Danh mục bánh Classic bánh chạy nhất với 14888 chiếc. Ba loại còn lại có sự chênh lệch không đáng kể với nhau, cụ thể Supreme là 11987, Veggie là 11649 và thấp nhất là Chicken với 11050 chiếc bánh.  
 Đối với từng danh mục pizza, doanh số không có sự chênh lệch nào đáng kể giữa các tháng.

**Kiểm tra sự chênh lệch trong doanh số giữa các danh mục pizza có ý nghĩa thống kê không bằng One-way ANOVA**

One-Way ANOVA (One-Way Analysis of Variance) là một phương pháp kiểm định thống kê được sử dụng để so sánh trung bình của nhiều nhóm dựa trên một đặc trưng. Phương pháp kiểm định này so sánh phương sai giữa các nhóm với phương sai trong từng nhóm thông qua F-statistic.





Kết quả có p-value (Pr(>F)) < 2e-16, gần bằng 0 nên ta có đủ bằng chứng để bác bỏ giả thuyết không. Bên cạnh đó, F value bằng 90, giá trị cao này cho thấy rằng sự khác biệt doanh số giữa các danh mục pizza lớn hơn nhiều so với sự biến động giữa các tháng của từng danh mục. Vì vậy, sự chênh lệch trong doanh số giữa các danh mục pizza là đáng kể hay nói cách khác là danh mục có tác động đến doanh số.

#### Phân tích doanh số bán hàng của từng bánh pizza

**Doanh số của từng bánh pizza trong cả năm 2015**

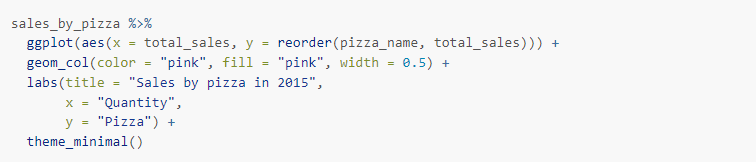
* Số bánh pizza cửa hàng kinh doanh

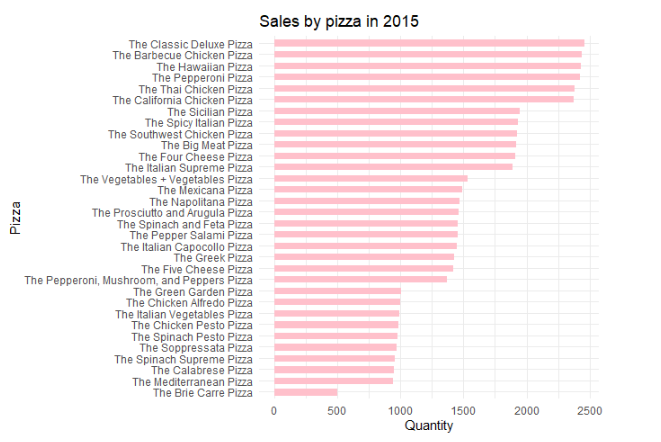




Cửa hàng kinh doanh 32 bánh pizza khác nhau.

* Trực quan hóa doanh số của từng bánh pizza trong một năm

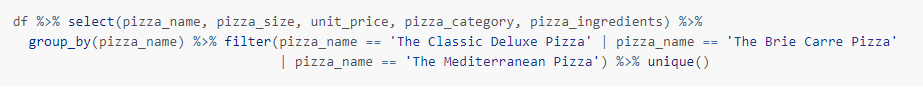


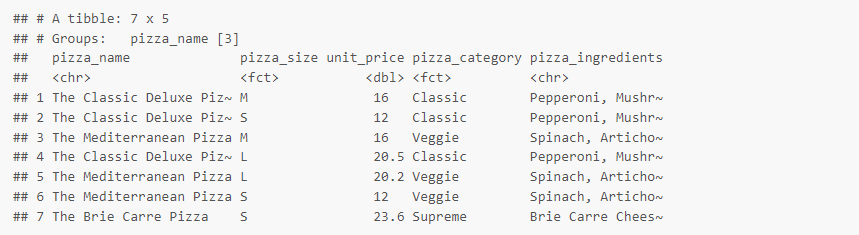


Pizza bán chạy nhất là The Classic Deluxe Pizza với 2453 bánh.

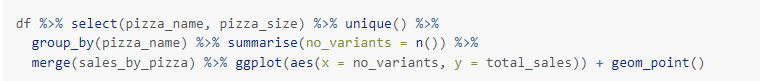
Pizza bán chậm nhất là The Brie Carre Pizza với 490 bánh. Ta có thể dễ dàng nhận thấy sự chênh lệch rất lớn giữa doanh số của nó với các bánh khác, trong đó, chỉ bằng khoảng 50% doanh số của pizza bán chậm thứ hai (The Mediterranean Pizza) và bằng khoảng 20% doanh số của bánh bán chạy nhất.

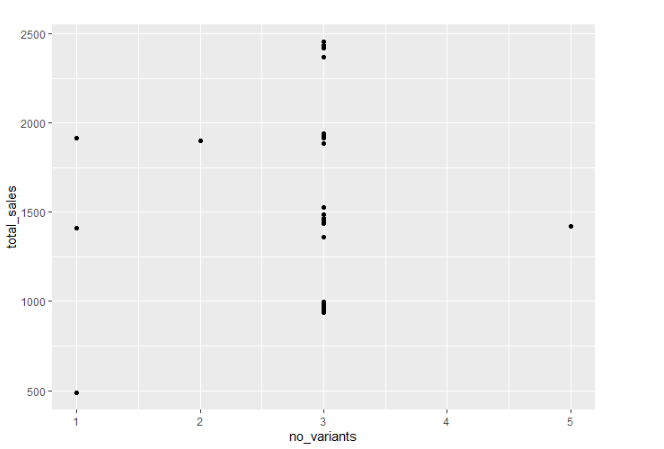
**Xem các thuộc tính của ba bánh này để dự đoán nguyên nhân gây ra sự chênh lệch lớn**





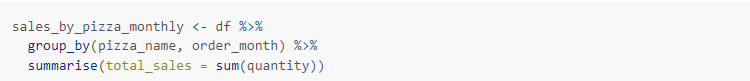
* Có thể nhận thấy sự chênh lệch rất lớn, gần gấp đôi trong giá của The Brie Carre Pizza size S (23.65 USD) và The Classic Deluxe Pizza hay The Mediterranean Pizza size S (12 USD). Do đó em cho rằng *biến unit\_price và pizza\_size có ảnh hưởng tới doanh số, hơn thế nữa hai biến này là hai biến tương tác*.
* Bên cạnh đó, sự khác biệt trong doanh số giữa The Classic Deluxe Pizza và The Mediterranean Pizza mặc dù chúng có mức giá gần như giống hệt nhau cho các kích cỡ. Điều này cho thấy *biến pizza\_category cũng ảnh hưởng tới doanh số*. Biến pizza\_ingredients sẽ trực tiếp phản ánh tới pizza\_cateogry nên em bỏ qua nó.
* Ngoài ra, The Brie Carre Pizza chỉ có một kích cỡ trong khi đa số bánh còn lại có nhiều kích cỡ, em nghi ngờ độ đa dạng của một bánh pizza ảnh hưởng tới doanh số của nó. Trực quan hóa để kiểm tra:

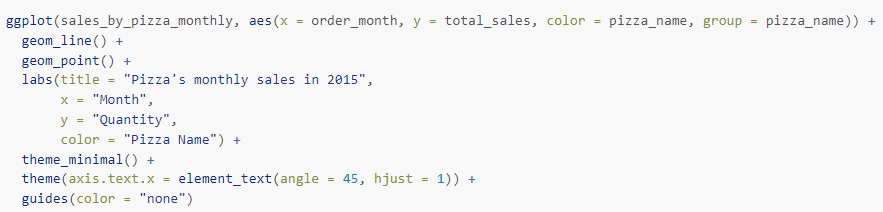


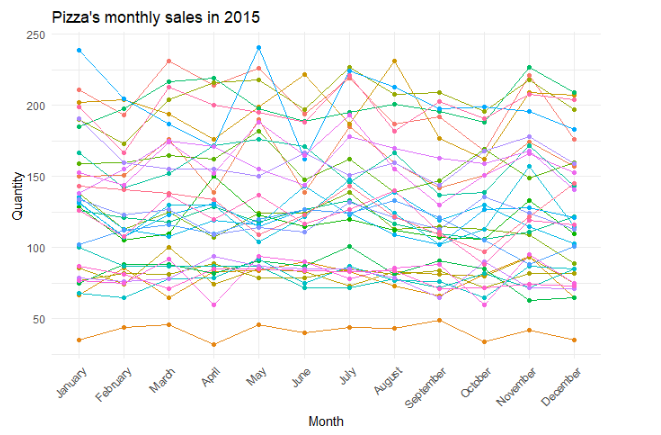


*Không có mối quan hệ tương quan giữa no\_variants và total\_sales, nghĩa là độ đa dạng của một bánh pizza không ảnh hưởng tới doanh số của nó.*

**Doanh số của từng bánh trong mỗi tháng**



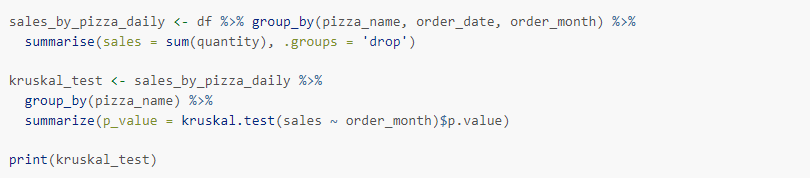


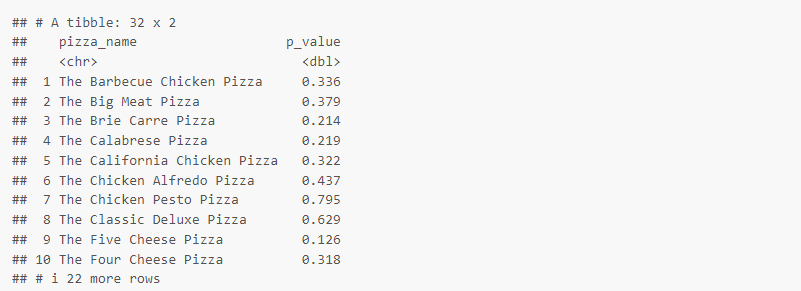


Mỗi đường trên biểu đồ thể hiện sự biến động trong doanh số giữa các tháng của một pizza. Có một số bánh có sự biến động trong doanh số một cách đáng chú ý.

**Kiểm định doanh số của từng pizza có khác biệt đáng kể giữa các tháng hay không bằng Kruskal-Wallis**

Kiểm định Kruskal-Wallis là một phương pháp thống kê phi tham số dùng để so sánh ba nhóm dữ liệu trở lên nhằm xác định xem có sự khác biệt đáng kể nào giữa các nhóm này hay không. Đây là một lựa chọn thay thế cho phân tích phương sai một chiều (ANOVA) khi dữ liệu không tuân theo phân phối chuẩn, phù hợp để kiểm tra sự biến động doanh số của một pizza qua các tháng có ý nghĩa hay không.  
Trong sales\_by\_pizza\_monthly, mỗi pizza chỉ có một quan sát cho một tháng nên kiểm định Kruskal-Wallis sẽ không có đủ thông tin để so sánh và phát hiện sự khác biệt có ý nghĩa. Do đó em sẽ tổng hợp doanh số của từng pizza theo ngày rồi thực hiện kiểm định cho từng pizza



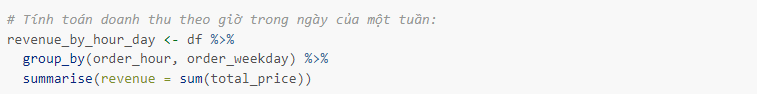


Kết quả không có p-value nào có giá trị nhỏ hơn 0.01, nên ta chấp nhận giả thuyết không và kết luận rằng không có sự khác biệt đáng kể trong doanh số giữa các tháng đối với mỗi pizza, hay nói cách khác, doanh số của mỗi pizza tương đối ổn định giữa các tháng trong năm 2015.

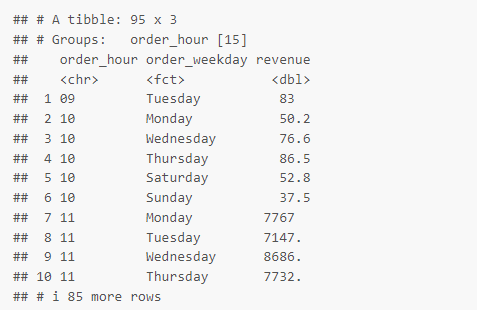
### Phân tích doanh thu theo thời gian

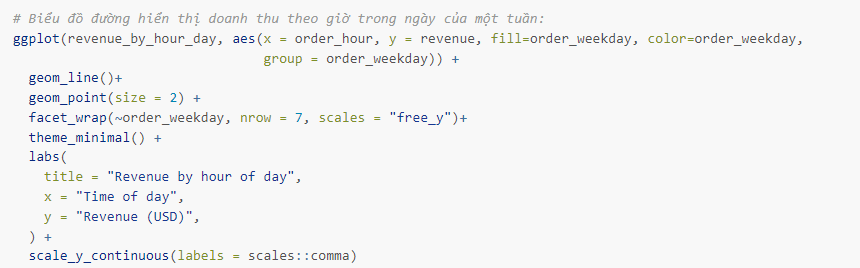
Phân tích tổng quan đã cho thấy trong năm 2015, cửa hàng có tổng doanh thu là 817,860 USD. Với giá trị đơn hàng nhỏ nhất là 9.75 USD, lớn nhất là 444.40 USD. Phần dưới đây, sẽ phân tích chi tiết hơn về doanh thu thu được của cửa hàng theo thời gian giờ trong ngày, tuần, và tháng trong năm 2015.

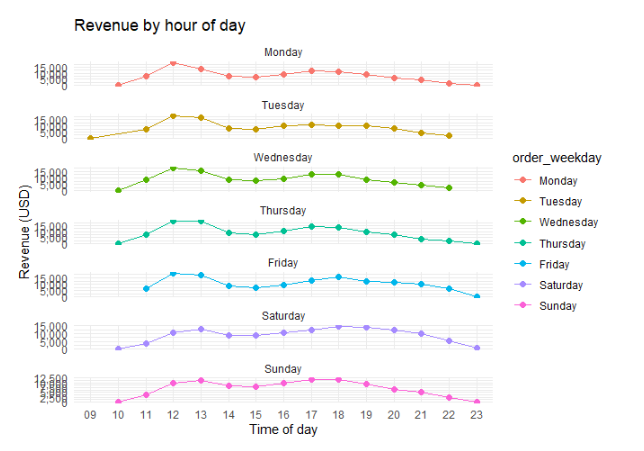
#### Phân tích doanh thu theo giờ trong ngày của một tuần











*Biểu đồ cho thấy doanh thu đều chạm đỉnh vào 12h các ngày trong tuần, trừ Thứ 7 và Chủ Nhật. Điều này cho thấy doanh thu sẽ cao vào giờ nghỉ trưa của những ngày đi làm của khách hàng. Ngày Thứ 7 và Chủ Nhật, doanh thu sẽ cao hơn vào thời điểm sau 17h đến 20h. Điều này có thể do khách hàng có xu hướng ăn tối muộn hơn vào cuối tuần. Bên cạnh đó, có sự chênh lệch doanh thu đáng kể vào các thời điểm trong ngày. Cửa hàng có thể triển khai các chương trình khuyến mãi vào các khung giờ thấp điểm để thu hút thêm khách hàng và tăng doanh thu.*

#### Phân tích doanh thu một ngày dùng Time Series – Chuỗi thời gian

1. Khái niệm về chuỗi thời gian:

Chuỗi thời gian là tập hợp các dữ liệu có sự phụ thuộc của các giá trị quan sát được thu thập qua thời gian. Với mục tiêu là dùng các dữ liệu trong quá khứ để phân tích và dự báo cho dữ liệu trong tương lai.

1. Các thành phần chuỗi thời gian:

* Thành phần xu hướng (Trend): Thành phần xu hướng là thành phần thể hiện mức tăng giảm của lượng biến quan sát theo một quy luật nào đó trong một khoảng thời gian đủ dài.
* Thành phần chu kỳ (Cyclic): Các mẫu chu kỳ xuất hiện dưới dạng những dao động không đều trong dữ liệu, bị ảnh hưởng bởi các điều kiện kinh tế và thường liên quan đến chu kỳ kinh doanh rộng lớn hơn. Khác với tính mùa vụ, ảnh hưởng chu kỳ đối với chuỗi thời gian trong tương lai là không thể đoán trước.
* Thành phần mùa (Seasonal): Trong khi thành phần xu hướng và thành phần chu kỳ của chuỗi thời gian được xác định bằng cách phân tích chuỗi thời gian kéo dài qua nhiều năm, thì nhiều chuỗi thời gian cho thấy có những kiểu biến thiên lặp lại theo tần suất theo quý , theo mùa,…, biến động của lượng biến tuân theo một quy luật nào đó giữa các thời điểm trong năm và lặp lại trong các năm kế tiếp.
* Thành phần bất thường: Ngoài các yếu tố xu hướng , chu kỳ, và mùa, thực tế giá trị lượng biến quan sát còn chịu sự chi phối của các yếu tố ngẫu nhiên, các yếu tố bất thường ngắn hạn không chịu sự liên kết với thành phần khác như mùa, chu kỳ, xu thế.

\*Phân biệt giữa các mẫu chu kỳ và mùa vụ:

- Các dao động không có tần suất cố định là chu kỳ.

- Nếu các dao động khớp với các khoảng thời gian theo lịch, chúng là mùa vụ.

- Nếu các dao động khớp với các khoảng thời gian theo lịch, chúng là mùa vụ.

Khi dữ liệu có yếu tố mùa, trước tiên cần tách yếu tố mùa ra khỏi chuỗi dữ liệu. Sau đó, sử dụng chuỗi dữ liệu được điều chỉnh yếu tố mùa để thực hiện dự báo xu thế. Có nhiều phương pháp để tách yếu tố mùa ra khỏi sữ liệu gốc theo nhiều phương pháp. Phương pháp đơn giản nhất là phương pháp trung bình trượt (Moving average method).

1. Đồ thị thời gian:

Đồ thị thời gian là những đồ thị biểu diễn các quan sát được vẽ theo các mốc thời gian tương ứng, với các điểm dữ liệu liên tiếp được nối bằng các đoạn thẳng.

1. Hệ số tự tương quan và nhận dạng dữ liệu:

Kiểm định sự bằng 0 của các hệ số tự tương quan:



Sử dụng các hệ số tự tương quan với các độ trễ khác nhau của một biến số nhất định Yt nhằm xác định tính chất của chuỗi:

* **Chuỗi ngẫu nhiên**: các giá trị kế nhau trong một chuỗi thời gian không có liên quan gì với nhau. ⇒ Hệ số tự tương quan giữa Yt và Yt−k với bất kỳ độ trễ k nào đều gần bằng 0.
* **Chuỗi dừng**: hệ số tự tương quan của các bậc đầu tiên khác 0 một cách có ý nghĩa thống kê, các hệ số tự tương quan các bậc sau bằng 0.
* **Chuỗi xu thế (không dừng):** các giá trị kế nhau trong một chuỗi thời gian có mức độ tương quan cao với nhau ⇒ Các hệ số tự tương quan khác 0 một cách có ý nghĩa thống kê.
* **Chuỗi có yếu tố mùa vụ:** hệ số tự tương quan tại một (một số) độ trễ mùa khác 0 một cách có ý nghĩa thống kê.

1. Chuỗi dừng:

Một số đặc điểm của chuỗi dừng:

* Thể hiện xu hướng trở lại trạng thái trung bình theo một cách trong đó dữ liệu dao động xung quanh một giá trị trung bình cố định trong dài hạn.
* Có một phương sai xác định không thay đổi theo thời gian.
* Có một giản đồ tự tương quan với các hệ số tương quan giảm dần khi độ trễ tăng.
* Lựa chọn mô hình dự báo

1. Lựa chọn mô hình phù hợp cho chuỗi thời gian:
2. Đối với dữ liệu dừng: sử dụng các mô hình dự báo thô, các phương pháp trung bình giản đơn, các mô hình trung bình di động, các mô hình ARIMA.
3. Đối với dữ liệu xu thế: sử dụng mô hình trung bình di động, san mũ Holt, hồi quy đơn, mô hình hàm xu thế, mô hình ARIMA.
4. Đối với dữ liệu mùa: sử dụng các mô hình phân tích, san mũ Winters, hồi quy bội, các mô hình ARIMA.
5. Đối với dữ liệu chu kỳ: sử dụng các mô hình phân tích, các mô hình kinh tế lượng, hồi quy bội, các mô hình ARIMA
6. Tiến hành phân tích:

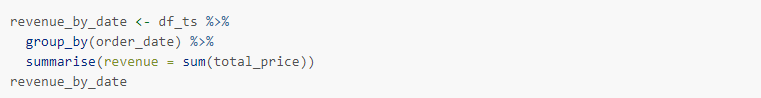
* Sao chép dữ liệu vào biến df\_ts để xử lý trên bản sao không gây ảnh hưởng đến dữ liệu gốc.



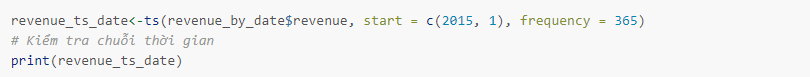
* Chuyển đổi cột order\_date thành kiểu Date

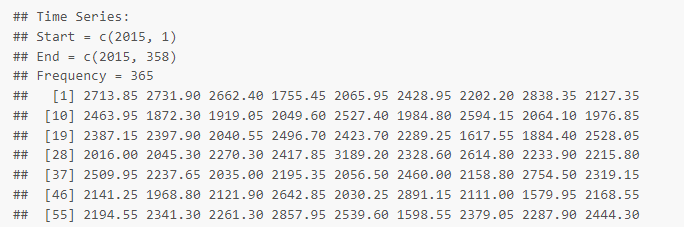


* Tính toán doanh thu theo ngày của một năm:



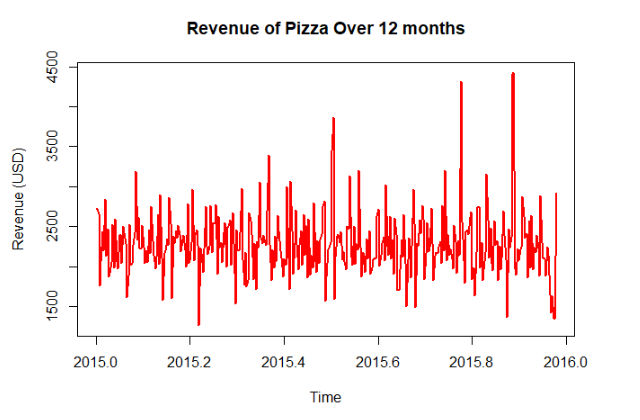
* Tạo chuỗi thời gian (time series) với tần suất 365 ngày (Năm 2015 không phải năm nhuận)





* Trực quan hóa doanh thu theo ngày với time\_series do có nhiều điểm dữ liệu:





*Xu hướng tổng thể: Có sự biến động trong doanh thu theo ngày trong năm 2015 của cửa hàng. Có những ngày doanh thu cao hơn và những ngày doanh thu thấp hơn. Nhưng xu hướng doanh thu tổng thể không tăng hoặc giảm rõ rệt trong giai đoạn từ 2015 đến 2016. Nó khá ổn định quanh mức trung bình. Doanh thu có sự biến động đáng kể hàng ngày, điều này có thể do nhiều yếu tố như ngày trong tuần, ngày lễ, hoặc các sự kiện đặc biệt. Xu hướng biến động: giảm vào đầu tháng và tăng vào cuối tháng. Doanh thu biến động ít hơn vào những ngày đầu năm. Doanh thu tăng cao bất thường vào những ngày của tháng 5, tháng 7 và khoảng tháng 8. Đầu năm 2015: Doanh thu dao động xung quanh mức trung bình, không có biến động đột biến. Giữa năm 2015: Xuất hiện một số đỉnh cao vượt xa mức trung bình, có thể do hoạt động khuyến mãi hoặc sự kiện đặc biệt. Cuối năm 2015: Có xu hướng giảm nhẹ, đặc biệt gần 2016.*

* Kiểm tra tính mùa vụ (Seasonality) của chuỗi thời gian:

Thời gian cần thiết:  
Tối thiểu một chu kỳ hoàn chỉnh để xác định mùa vụ. Tối thiểu hai chu kỳ để xác nhận tính lặp lại và phân tích chính xác hơn.  
Ví dụ:  
Chu kỳ hàng tuần (7 ngày): Cần ít nhất 7-14 ngày.  
Chu kỳ hàng tháng (12 tháng): Cần ít nhất 12-24 tháng.  
Chu kỳ hàng năm (365 ngày): Cần ít nhất 365-730 ngày (1-2 năm).

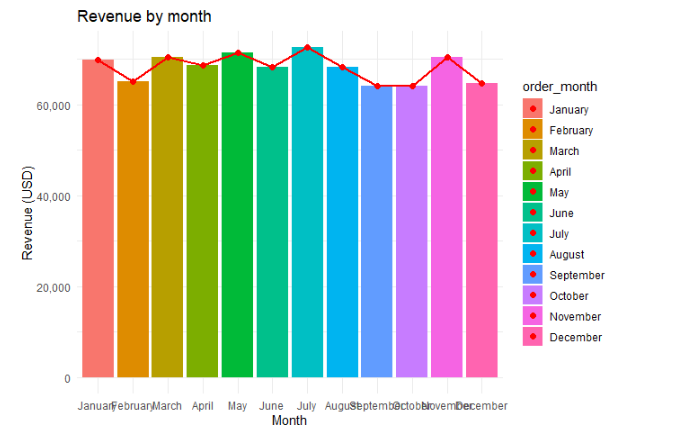
*Do dữ liệu chỉ các ngày trong một năm 2015 nên chuỗi thời gian revenue\_ts\_date không có tính mùa vụ.*

#### Phân tích doanh thu theo tháng



* Trực quan hóa bằng biểu đồ cột để so sánh doanh thu giữa các tháng trong năm:





Trong năm 2015, doanh thu cao nhất được ghi nhận vào tháng 7 với hơn 70.000 USD. Các tháng có nhu cầu cao nhất là Tháng Bảy và Tháng Mười Một, nhưng Tháng Mười Một lại có doanh thu thấp hơn so với Tháng Năm. Điều này cho thấy rằng dù có nhu cầu tiêu thụ cao vào tháng này, nhưng doanh thu không tương ứng, có thể do các chương trình khuyến mãi hoặc giá bán không phù hợp.

Biểu đồ đường cho thấy doanh thu giữa các tháng về cơ bản không có sự chênh lệch lớn, doanh thu ổn định ở mức trên 60.000 USD mỗi tháng. Điều này cho thấy cửa hàng duy trì được sự ổn định trong doanh thu hàng tháng, tuy nhiên, cần có các chiến lược cụ thể để tối ưu hóa doanh thu trong các tháng có nhu cầu cao.

## Xây dựng mô hình dự đoán

### Dự đoán doanh số theo sản phẩm

Mục tiêu của phần này là xây dựng mô hình dự đoán doanh số trong một tháng của một bánh pizza cụ thể, tức là dự đoán xem cửa hàng bán được bao nhiêu bánh pizza đó bằng các đặc trưng của nó cùng với dữ liệu bán hàng trong quá khứ. Em lựa chọn sử dụng hai mô hình là Linear Regression và Random Forest nhằm tận dụng những ưu điểm riêng của mỗi phương pháp và đánh giá hiệu quả của chúng trong bối cảnh dữ liệu hiện có.

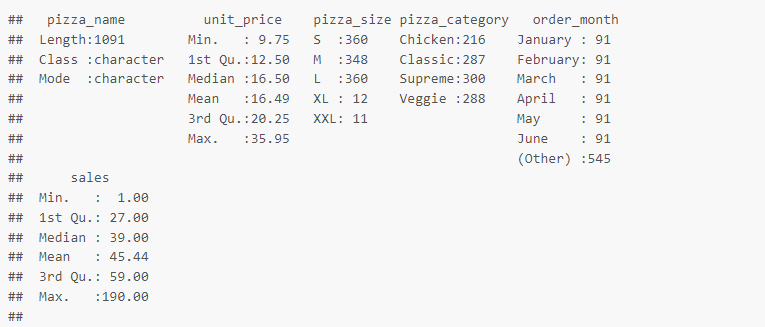
* Tổng hợp dữ liệu chứa doanh số mỗi tháng của từng bánh pizza









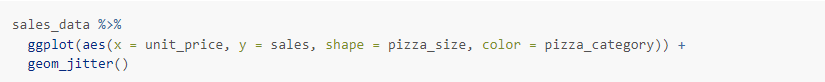


Dữ liệu gồm 1091 quan sát, mỗi quan sát thể hiện doanh số của một bánh với kích cỡ xác định (hoặc S hoặc M hoặc L hoặc XL hoặc XXL), đơn giá tương ứng, thuộc danh mục nào và doanh số trong một tháng cụ thể là bao nhiêu.

Biến sales là biến phụ thuộc mà ta xây dựng mô hình để dự đoán, dao động từ 1 đến 444, với trung bình là 115, đây sẽ là cơ sở để đánh giá mô hình.

Các biến độc lập bao gồm: pizza\_name, unit\_price, pizza\_size và pizza\_category. Như đã kiểm định khi phân tích doanh số của từng pizza (6.3.3), doanh số giữa các tháng của một bánh pizza tương đối ổn định, nên không chọn order\_month làm biến độc lập.

* Trực quan hóa mối quan hệ giữa doanh số và giá cùng với kích cỡ và danh mục bánh.



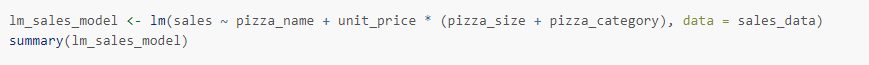


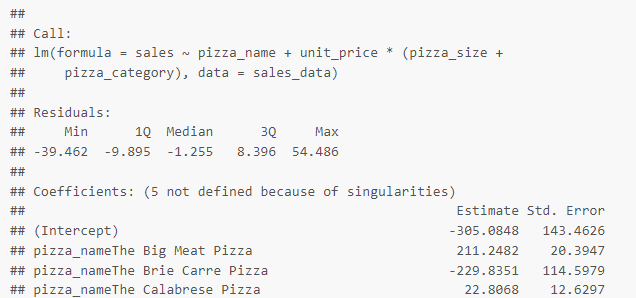
Biểu đồ cho thấy tác động của giá bán lên sales khác nhau đối với các cỡ bánh khác nhau và các loại bánh khác nhau. Cụ thể như bánh size S giá thấp (~10) có doanh số cao hơn rất nhiều so với bánh size S giá trên 20 USD. Và cùng ở khoảng giá 20 USD nhưng doanh số của các danh mục chênh lệch nhau rất rõ rệt.

*Vì vậy, ngoài 4 biến độc lập đã nêu (pizza\_name, unit\_price, pizza\_size và pizza\_category) thì em sẽ thêm sự tác giữa unit\_price và pizza\_size, giữa unit\_price và pizza\_category vào mô hình dự đoán.*

#### Dự đoán doanh số của sản phẩm bằng hồi quy tuyến tính

Hồi quy tuyến tính (Linear Regression) là mô hình thống kê đơn giản và dễ hiểu, dựa trên giả định về mối quan hệ tuyến tính giữa biến độc lập (biến đầu vào) và biến phụ thuộc (biến mục tiêu). Lý do em chọn mô hình này để dự đoán doanh số của sản phẩm vì nó đơn giản và dễ diễn giải, giúp em bước đầu hiểu rõ mức độ và hướng ảnh hưởng của từng biến độc lập đến doanh số sản phẩm. Hơn nữa, hồi quy tuyến tính là một mô hình nền tảng và thường được dùng như tiêu chuẩn so sánh cho các mô hình phức tạp hơn. Khi dữ liệu có dạng tuyến tính hoặc gần tuyến tính, Linear Regression cũng có khả năng mang lại kết quả dự đoán chính xác với chi phí tính toán thấp.



 F-statistic gần bằng 56 với p-value < 2.2e-16, cho thấy rằng **mô hình hồi quy này có ý nghĩa thống kê**. Cụ thể về ý nghĩa thống kê của từng biến, các biến giả của pizza\_category bị thiếu giá trị (NA), nhưng các biến tương tác của nó với unit\_price có p-value rất thấp, chứng tỏ **đưa sự tương tác này vào mô hình là hợp lý**. Một số biến dummy của pizza\_name, pizza\_size có p-value lớn hơn 0.05 nhưng đa số có p-value nhỏ, chứng tỏ **các biến độc lập này có ý nghĩa trong mô hình**.

Multiple R-squared bằng 0.6862 và Adjusted R-squared bằng 0.6739 cho **thấy mô hình giải thích khoảng 67% sự biến động của doanh số trong tập dữ liệu**. Mặc dù không quá cao, giá trị này vẫn cho thấy mô hình có thể giải thích một phần đáng kể sự biến động doanh số trong tập dữ liệu.

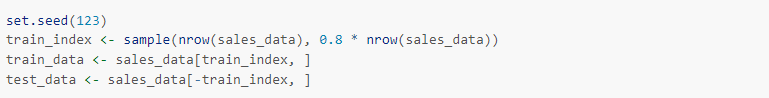
#### Dự đoán doanh số theo sản phẩm bằng Random Forest

Random Forest là một mô hình học máy có giám sát (supervised learning), bao gồm nhiều cây quyết định (decison tree), thường được dùng khi các quan hệ giữa các biến đầu vào và biến mục tiêu là phức tạp hoặc không tuyến tính. Em chọn Random Forest để so sánh với Linear Regression, vì mô hình này có khả năng xử lý và học những mối quan hệ phức tạp giữa các yếu tố ảnh hưởng đến doanh số của sản phẩm như đơn giá, kích cỡ,… mà mô hình tuyến tính không thể nắm bắt.

* Mỗi cây trong rừng được xây dựng theo các bước:
* Tạo tập dữ liệu huấn luyện: Từ tập dữ liệu huấn luyện ban đầu (train data), thuật toán tạo ra tập dữ liệu huấn luyện mới bằng cách lấy mẫu ngẫu nhiên có hoàn lại (bootstrap sampling with replacement) trên 2/3 dữ liệu. 1/3 dữ liệu bị bỏ qua được gọi là out-of-bag samples.
* Huấn luyện cây: Tại mỗi bước tách nhánh (node split), Random Forest không sử dụng toàn bộ các đặc trưng mà chỉ chọn ngẫu nhiên m đặc trưng (mtry). Điều này giúp giảm sự phụ thuộc vào bất kỳ đặc trưng nào và tăng tính đa dạng của các cây.
* Kết quả cuối cùng của bài toán được tính bằng cách lấy trung bình kết quả của tất cả các cây trong rừng (do đây là bài toán hồi quy, đối với bài toán phân loại thì lấy đa số phiếu).

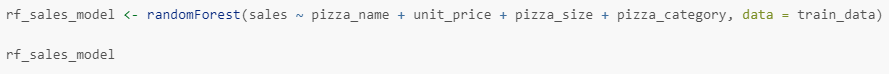
**Thực hiện xây dựng mô hình:**

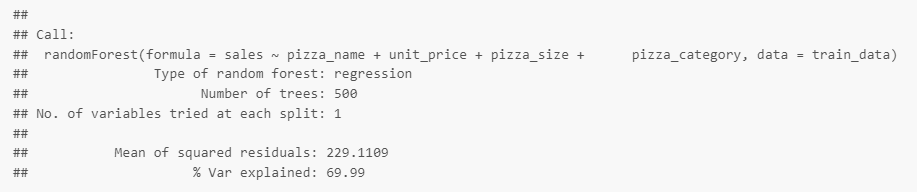
* Chia tập dữ liệu thành 80% cho huấn luyện và 20% cho kiểm tra.



* Xây dựng mô hình dự đoán doanh số

Trong công thức của mô hình này, em không chỉ định cụ thể sự tương tác giữa các biến độc lập như trong mô hình hồi quy tuyến tính vì random forest có khả năng tự học và nắm bắt các tương tác phức tạp giữa các biến.





Thông tin về mô hình:

* Đây là một mô hình random forest cho bài toán hồi quy, tức là nó dự đoán giá trị liên tục, trong bài toán này là doanh số của sản phẩm.
* Mô hình bao gồm 500 cây quyết định. Số cây càng lớn thì mô hình càng trở nên ổn định và chính xác hơn, nhưng cũng sẽ tốn thời gian huấn luyện và yêu cầu bộ nhớ cao hơn.
* Số biến được thử ở mỗi lẫn phân tách là 1, nghĩa là với mỗi lần phân chia trong các cây, thuật toán sẽ ngẫu nhiên chọn 1 biến (trong tổng số 3 biến: unit\_price, pizza\_size, và pizza\_category) và tính điểm phân chia tốt nhất dựa trên chúng.
* Từ giá trị trung bình bình phương phần dư (Mean of squared residuals), ta có sai lệch trung bình giữa các giá trị dự đoán và giá trị thực tế của doanh số là khoảng 15 đơn vị (RMSE), bằng 13% giá trị trung bình của sales.
* Mô hình đã giải thích được 69% sự biến động của doanh số dựa trên các đặc trưng đầu vào.

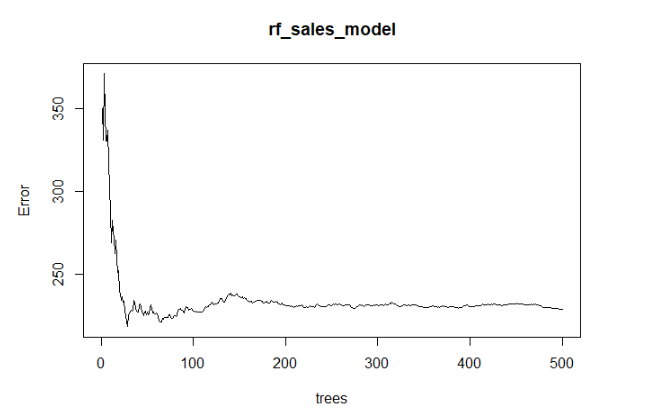
*Mô hình này khá đáng tin cậy (RMSE khá thấp) nhưng khả năng dự đoán vẫn còn khá thấp.*

**Tối ưu hóa mô hình**

Hiện tại, số cây quyết định của mô hình là 500 và số biến được thử tại mỗi lần phân tách là 1.

* Kiểm tra ngưỡng cây tối ưu chưa:

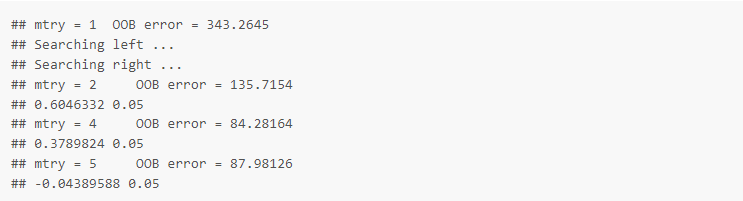


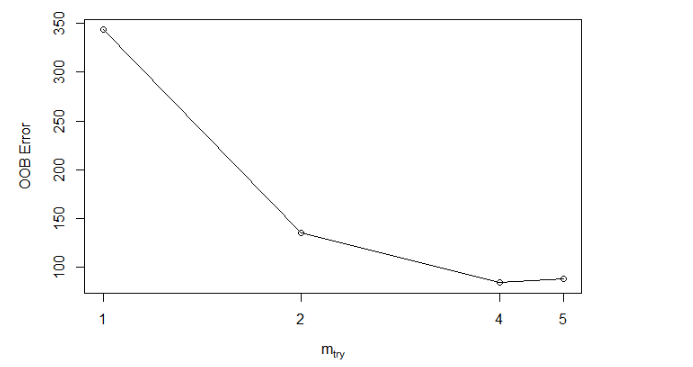


Biểu đồ cho thấy **500 cây là phù hợp** vì error tại đây đã rất nhỏ và gần như không còn giảm nữa.

* Sử dụng tuneRF để tìm tham số mtry tối ưu cho mô hình:

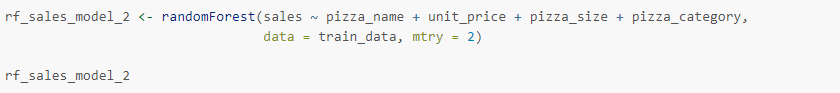


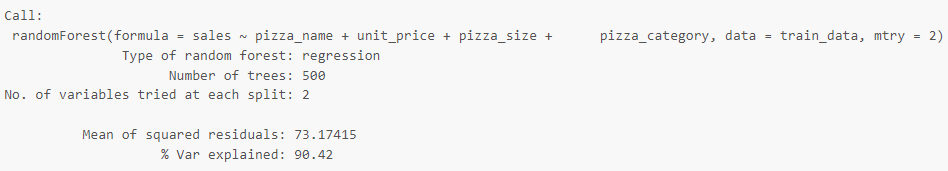




OOB error rate là lỗi ngoài bộ dữ liệu huấn luyện, được tính dựa trên sự chênh lệch giữa giá trị thực và giá trị dự đoán của các mẫu OOB. Thông số này phản ánh độ chính xác của mô hình, OOB error rate càng thấp thì mô hình càng chính xác. Dự vào biểu đồ kết hợp với việc dữ liệu có 4 biến độc lập, mà nhiều biến có sự tương tác phức tạp với nhau nên em **chọn mtry = 2 và 3 để tối ưu mô hình.**

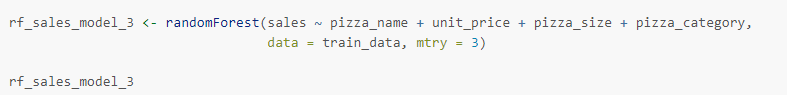
* Thực hiện tối ưu mô hình với mtry = 2:

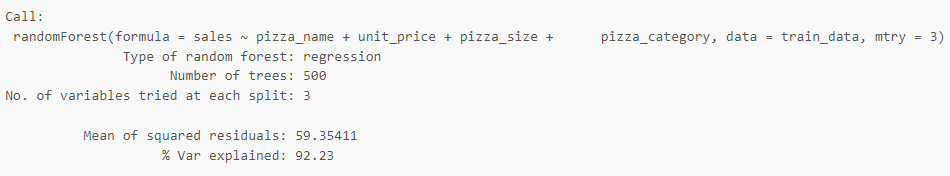




*Từ Mean of squard residuals ta có RMSE bằng khoảng 8.6, giảm 6 đơn vị so với mô hình random forest ban đầu. Và % Var explained bằng 90%, tăng 20%. Chứng tỏ giá trị mtry = 2 là giá trị phù hợp để tối ưu mô hình.*

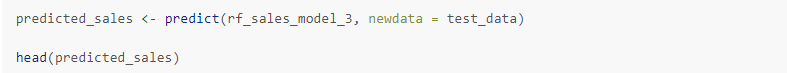
* Thực hiện tối ưu mô hình với mtry = 3:





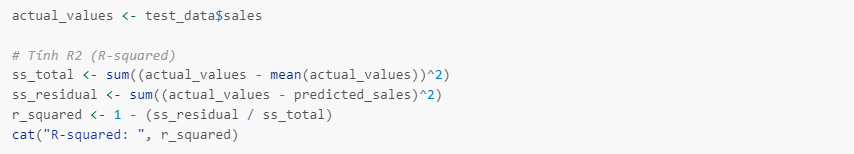
*Kết quả cho thấy Mean of squard residuals thấp và % Var explained cao hơn cả mô hình có m=2. Chọn đây là mô hình dự đoán cuối cùng.*

* Dự đoán doanh số trên tập dữ liệu kiểm tra





* Kiểm tra độ chính xác





Mô hình đã giải thích được 92% sự biến thiên của sales trong tập dữ liệu kiểm tra.

* Tính toán lỗi RMSE (Root Mean Squared Error)





*Mô hình có R-squared cao trên cả tập dữ liệu huấn luyện và tập kiểm tra, nên có thể kết luận rằng mô hình không bị overfitting và đạt hiệu quả dự đoán rất tốt. Bên cạnh đó, sai số dự đoán (RMSE) nhỏ cho thấy đây là mô hình rất đáng tin cậy để dự đoán doanh số.*

#### Dự đoán doanh số bằng mô hình GBM

Boosting là một phương pháp mạnh mẽ trong học máy, được thiết kế để kết hợp nhiều mô hình yếu (weak learners) thành một mô hình mạnh (strong learner). Bằng cách xây dựng tuần tự các mô hình và tập trung vào các mẫu dữ liệu khó dự đoán, boosting có thể cải thiện độ chính xác và khả năng tổng quát hóa của mô hình.

1. Boosting: Tổng quan và cơ chế hoạt động

Boosting hoạt động dựa trên nguyên tắc:

* **Kết hợp các mô hình yếu:** Thay vì xây dựng một mô hình phức tạp từ đầu, boosting tạo ra nhiều mô hình đơn giản, mỗi mô hình tập trung vào việc sửa lỗi của mô hình trước đó.
* **Tối ưu hóa tổng quát:** Các mô hình được kết hợp bằng cách gán trọng số cho từng mô hình dựa trên hiệu suất của chúng.

1. Gradient và vai trò trong Boosting

Gradient trong boosting biểu thị độ dốc của hàm lỗi đối với từng mẫu dữ liệu. Trong quá trình xây dựng mô hình, boosting sử dụng gradient để:

* **Học từ lỗi:** Tăng cường khả năng dự đoán của mô hình bằng cách tập trung vào các mẫu dữ liệu khó.
* **Cập nhật mô hình:** Gradient được sử dụng để tối ưu hóa các cây quyết định trong từng bước lặp.

1. GBM:

**3.1. Khái niệm và ưu điểm**

Gradient Boosting Machine (GBM) là một thuật toán học máy mạnh mẽ, dựa trên nguyên tắc boosting, được thiết kế để tối ưu hóa hiệu suất dự đoán qua việc xây dựng tuần tự các mô hình yếu và sử dụng gradient để cập nhật mô hình. GBM nổi bật với những ưu điểm sau:

* **Hiệu suất vượt trội:** GBM có khả năng xử lý tốt các bài toán hồi quy và phân loại, đặc biệt trong các tình huống dữ liệu phức tạp với mối quan hệ phi tuyến.
* **Khả năng tối ưu hóa linh hoạt:** GBM cho phép lựa chọn nhiều hàm mất mát khác nhau (như MSE, Log Loss), giúp thích nghi với các bài toán đa dạng.
* **Kiểm soát hiện tượng quá khớp:** Bằng cách điều chỉnh các tham số như learning rate, max\_depth và n\_estimators, GBM giúp mô hình cân bằng giữa độ chính xác và khả năng tổng quát hóa.

**3.2. Các tham số quan trọng trong GBM**

Một số tham số quan trọng của GBM bao gồm

* **n\_estimators:** Số lượng cây quyết định (decision trees) trong mô hình. Mỗi cây sẽ được thêm vào dần dần để cải thiện mô hình. Giá trị này càng lớn sẽ làm mô hình mạnh mẽ hơn, nhưng cũng dễ dẫn đến overfitting.
* **learning\_rate:** Tốc độ học của mô hình, điều chỉnh ảnh hưởng của mỗi cây đối với kết quả cuối cùng. Nếu giá trị này quá lớn, mô hình có thể học quá nhanh và không tối ưu, còn nếu quá nhỏ, mô hình sẽ học chậm và có thể cần nhiều cây hơn.
* **max\_depth:** Độ sâu tối đa của mỗi cây quyết định. Cây càng sâu thì có thể học được nhiều mối quan hệ phức tạp hơn, nhưng nếu quá sâu có thể dẫn đến overfitting.
* min\_samples\_split: Số mẫu tối thiểu để chia một node. Đây là tham số điều chỉnh

1. Dự đoán bằng mô hình GBM

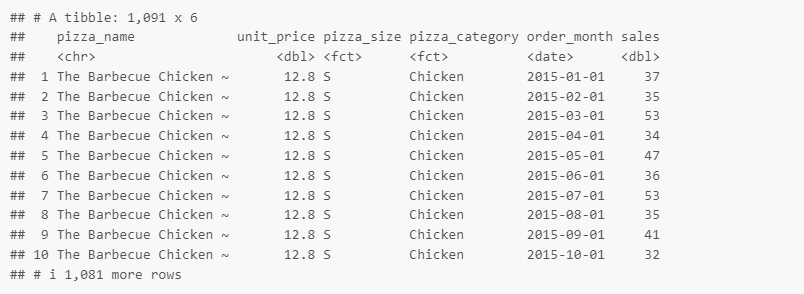
Ở mô hình Random Forest xây dựng các cây quyết định độc lập song song và kết hợp chúng bằng trung bình hoặc biểu quyết đa số, dẫn đến mô hình ít nhạy cảm hơn với lỗi nhưng cũng có thể bỏ qua các mẫu quan trọng. Boosting, với khả năng học có trọng số, có xu hướng phù hợp hơn để nắm bắt các mối quan hệ phi tuyến phức tạp giữa các đặc trưng và mục tiêu, đặc biệt khi dữ liệu có nhiễu hoặc cần tối ưu hóa sâu hơn.

* Tại sao sử dụng mô hình GBM:

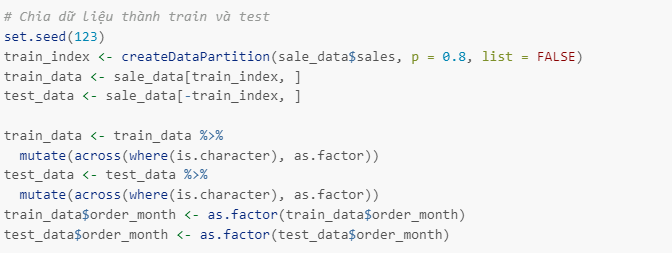
Với bài toán dự đoán doanh số bán hàng pizza, GBM là một lựa chọn hợp lý do dữ liệu có nhiều đặc trưng dạng danh mục và mối quan hệ giữa các đặc trưng không hoàn toàn tuyến tính. Khả năng sử dụng các tham số tùy chỉnh như độ sâu cây, tốc độ học, và số cây cũng cho phép chúng em tinh chỉnh mô hình để phù hợp với cấu trúc của dữ liệu này.

* Tạo dataframe phù hợp và thực hiện tiền xử lý để phù hợp với yêu cầu của mô hình GBM.

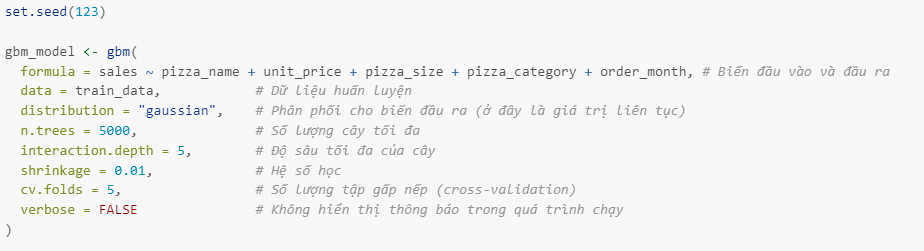




* Chia dữ liệu thành train và test



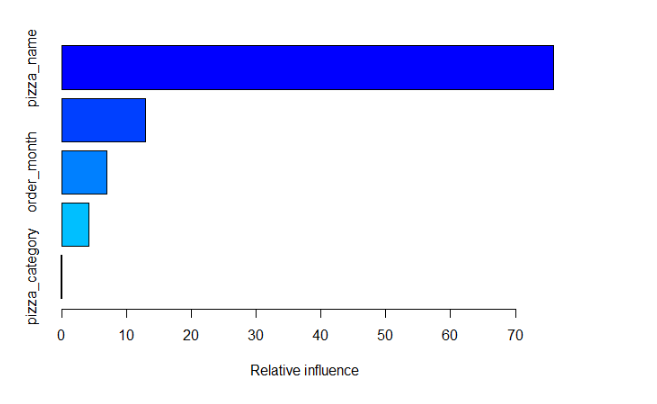
* Huấn luyện mô hình GBM

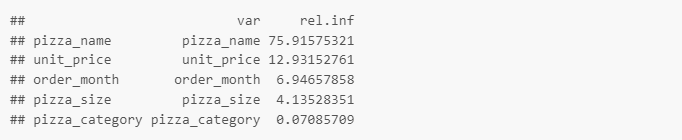


Chúng em sử dụng phân phối “gaussian” cho biến phụ thuộc trong mô hình này vì “gaussian” được sử dụng khi biến đầu ra là một giá trị liên tục (ví dụ, giá trị dự đoán cho sales là một số thực). Đây là lựa chọn chuẩn cho các bài toán hồi quy.

* Xem thông tin về relative influence của các đặc trưng:



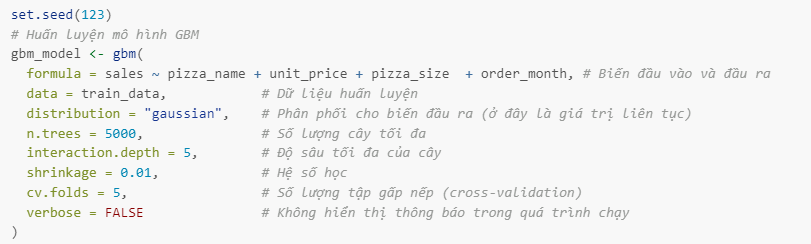




Từ kết quả của summary(gbm\_model) cho thấy giá trị Relative Influence của đặc trưng pizza\_category chỉ đạt 0.03, cho thấy đặc trưng này gần như không đóng góp đáng kể vào quá trình dự đoán doanh số bán hàng (sales) so với các đặc trưng khác. Điều này có thể ngụ ý rằng sự phân loại theo danh mục pizza (pizza\_category) không mang lại thông tin hữu ích hoặc không tương quan mạnh với biến mục tiêu trong dữ liệu huấn luyện. Việc giữ lại một đặc trưng có tầm quan trọng thấp như vậy có thể làm tăng độ phức tạp không cần thiết của mô hình mà không cải thiện đáng kể hiệu suất dự đoán. Do đó, trong các bước tối ưu hóa, pizza\_category sẽ được loại bỏ để giảm độ phức tạp và tăng hiệu quả tính toán mà vẫn đảm bảo độ chính xác của mô hình.

**Tối ưu hóa mô hình**

* Loại bỏ biến độc lập pizza\_category

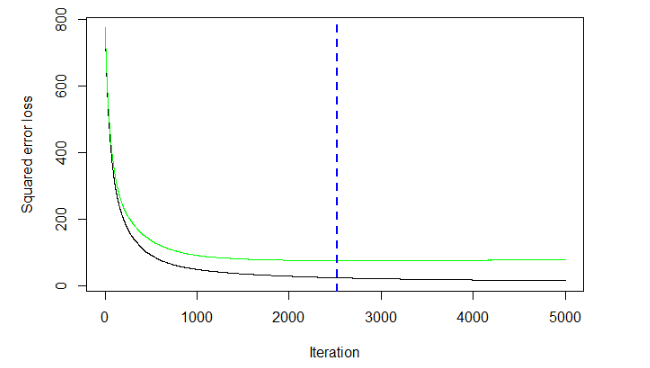


* Xác định số lượng cây tối ưu dựa trên kĩ thuật cross-validation

Việc chọn số lượng cây tối ưu rất quan trọng, vì nếu số lượng cây quá ít, mô hình sẽ chưa đủ mạnh để học các mẫu dữ liệu phức tạp; nếu quá nhiều cây, mô hình có thể trở nên quá phức tạp và dễ bị overfitting.

Em sẽ dựa vào Cross-validation loss để tính số lượng cây tối ưu vì nó phản ánh khả năng tổng quát của mô hình đối với dữ liệu chưa được huấn luyện, giúp đánh giá mô hình một cách khách quan và tránh overfitting.



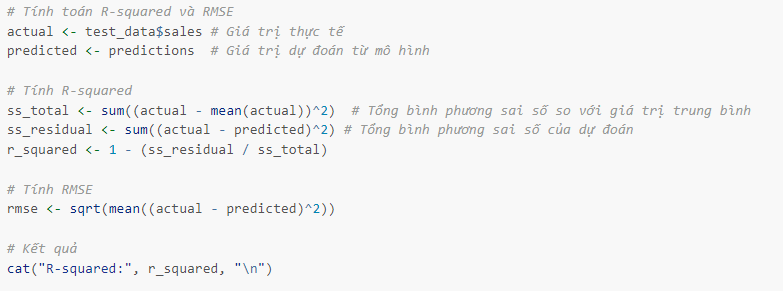


Sai số trên tập huấn luyện (Training Loss) giảm đều khi tăng số lượng cây, điều này cho thấy mô hình dần học tốt hơn trên tập huấn luyện.  
Sai số kiểm định chéo (CV Loss) giảm đến một mức tối thiểu rồi có xu hướng tăng trở lại, cho thấy điểm này là số cây tối ưu để tránh overfitting.

* Dự đoán trên tập kiểm tra với best trees



* Kiểm tra tính chính xác của mô hình GBM

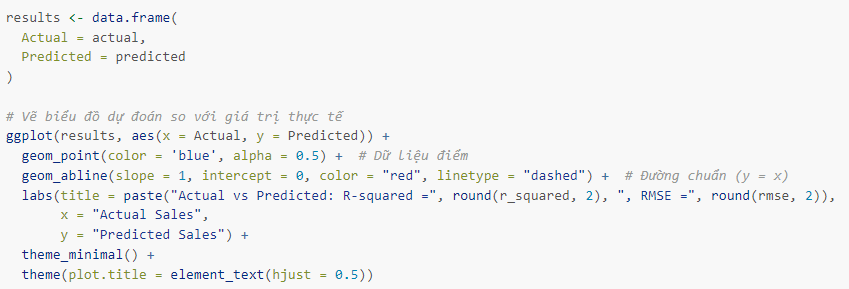


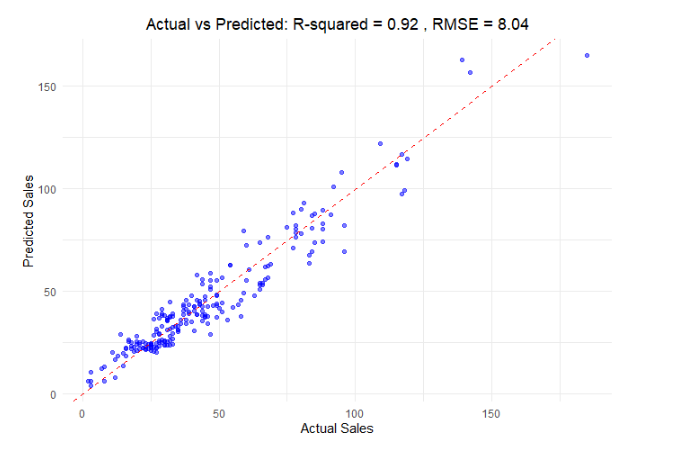






* Trực quan hóa độ chính xác trong mô hình





Mô hình Gradient Boosting Machine (GBM) dự đoán doanh số bán pizza đã đạt được hiệu quả cao, thể hiện qua giá trị R-squared là 0.92, cho thấy mô hình giải thích được 92% biến thiên của dữ liệu. Giá trị Root Mean Squared Error (RMSE) là 8.04, cho thấy sai số trung bình giữa giá trị dự đoán và giá trị thực tế là khoảng 8 đơn vị. Kết quả này chứng tỏ mô hình có độ chính xác rất tốt.

#### Kết luận

Trong bài toán này, mô hình hồi quy tuyến tính có độ chính xác thấp nhất (~67%), trong khi hai mô hình Random forest và GBM đều cho sai số dự đoán nhỏ tương đương nhau và đều có độ chính xác cao khoảng 92%. Vì vậy, em chọn random forest là mô hình cuối cùng dùng để dự đoán doanh số của sản phẩm. Quyết định này là do khi hai mô hình có hiệu suất tương đương thì nên chọn mô hình đơn giản hơn để dễ hiểu, bảo trì, triển khai, ít nguy cơ overfitting và yêu cầu ít tài nguyên tính toán hơn. Trong bài toán này, Random forest có khả năng tổng quát hóa tốt hơn mà không cần điều chỉnh quá nhiều, và các cây quyết định của nó được xây dựng một cách độc lập với nhau nên có tốc độ huấn luyện nhanh và ít nhạy cảm với nhiễu trong tập dữ liệu hơn mô hình GBM.  
Ngoài ra, mô hình random forest sau khi được tối ưu đã dự đoán doanh số của sản phẩm tốt hơn rất nhiều so với mô hình hồi quy tuyến tính vì nó có khả năng nắm bắt các mối quan hệ phi tuyến tính và tương tác phức tạp giữa các biến độc lập mà mô hình hồi quy tuyến tính chưa mô tả được. Mô hình random forest còn có khả năng xử lý hiệu quả biến phân loại nhiều mức một cách tự nhiên khi xây dựng các cây quyết định, trong khi đó mô hình hồi quy tuyến tính sẽ dùng kỹ thuật one hot encoding để tạo biến giả, tăng độ phức tạp của mô hình và có thể gây ra vấn đề đa cộng tuyến. Hơn nữa, random forest ít bị ảnh hưởng bởi giá trị ngoại lai (outliers) nhờ lấy trung bình của nhiều cây quyết định, và không yêu cầu các giả định phân phối nghiêm ngặt, phù hợp với tính chất dữ liệu trong bài toán này.

### Dự đoán doanh thu trong 6 tháng tiếp theo bằng mô hình ARIMA

1. **Giới thiệu về mô hình ARIMA:**

**Mô hình ARIMA**: Dựa trên giả thuyết chuỗi dừng và phương sai sai số không đổi. Mô hình sử dụng đầu vào chính là những tín hiệu quá khứ của chuỗi được dự báo để dự báo nó. Các tín hiệu đó bao gồm: chuỗi tự hồi qui AR (Auto Regression) và chuỗi trung bình trượt MA (Moving Average). Hầu hết các chuỗi thời gian sẽ có xu hướng tăng hoặc giảm theo thời gian, do đó yếu tố chuỗi dừng thường không đạt được. Trong trường hợp chuỗi không dừng thì ta sẽ cần biến đổi sang chuỗi dừng bằng sai phân. Khi đó tham số đặc trưng của mô hình sẽ có thêm thành phần bậc của sai phân d và mô hình được đặc tả bởi 3 tham số ARIMA(p, d, q).

1. **Lý thuyết về mô hình ARIMA:**

ARIMA model là viết tắt của cụm từ Autoregressive Intergrated Moving Average, đây là một kỹ thuật được sử dụng để phân tích chuỗi thời gian và dự báo các giá trị tương lai có thể của chuỗi thời gian. Mô hình tự hồi quy (**Autoregressive modeling**) và mô hình trung bình trượt (**Moving Average modeling**) là hai phương pháp khác nhau để dự báo dữ liệu chuỗi thời gian. ARIMA tích hợp cả hai phương pháp này, do đó có tên gọi như vậy.

ARIMA là một trong những phương pháp được sử dụng rộng rãi nhất trong dự báo chuỗi thời gian và có thể được sử dụng theo hai cách khác nhau tùy thuộc vào loại dữ liệu chuỗi thời gian.

Trong trường hợp đầu tiên, chúng ta tạo ra một mô hình ARIMA không mùa vụ (Non-seasonal ARIMA) mà không cần phải tính đến tính mùa vụ trong dữ liệu chuỗi thời gian. Chúng ta dự báo tương lai đơn giản dựa trên các mẫu trong dữ liệu quá khứ.

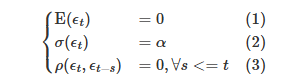
Trong trường hợp thứ hai, chúng ta xem xét mùa vụ, tức là các chu kỳ định kỳ ảnh hưởng đến chuỗi thời gian. Những chu kỳ này có thể là hàng ngày, hàng tuần hoặc hàng tháng và giúp xác định các mẫu trong dữ liệu quá khứ của chuỗi thời gian, từ đó có thể dùng để dự báo giá trị tương lai.

Mô hình sẽ biểu diễn phương trình hồi qui tuyến tính đa biến (Multiple Linear Regression) của các biến đầu vào (còn gọi là biến phụ thuộc trong thống kê) là 2 thành phần chính:

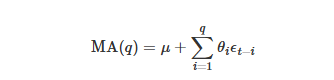
**Auto regression**: Kí hiệu là AR. Đây là thành phần tự hồi quy bao gồm tợp hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể, quá trình AR(p) của chuỗi xt được biểu diễn như bên dưới:



**Moving average:** Quá trình trung bình trượt được hiểu là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuổi theo thời gian. Do chuỗi của chúng ta được giả định là dừng nên quá trình thay đổi trung bình dường như là một chuỗi nhiễu trắng. Quá trình moving average sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên. Chuỗi này phải là một chuỗi nhiễu trắng thỏa mãn các tính chất:



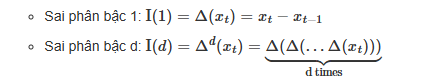
Vế (1) có nghĩa rằng kì vọng của chuỗi bằng 0 để đảm bảo chuỗi dừng không có sự thay đổi về trung bình theo thời gian. Vế (2) là phương sai của chuỗi không đổi. Do kì vọng và phương sai không đổi nên chúng ta gọi phân phối của nhiễu trắng là phân phối xác định (identical distribution) và được kí hiệu là ϵt∼WN(0,σ2). Nhiễu trắng là một thành phần ngẫu nhiên thể hiện cho yếu tố không thể dự báo của model và không có tính qui luật. Quá trình trung bình trượt được biểu diễn theo nhiễu trắng như sau:



Quá trình này có thể được biểu diễn theo [**dịch chuyển trễ - backshift operator**](https://en.wikipedia.org/wiki/Lag_operator) B như sau:



**Intergrated**: Là quá trình đồng tích hợp hoặc lấy sai phân. Yêu cầu chung của các thuật toán trong Time Series là chuỗi phải đảm bảo tính dừng. Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc là thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là chúng ta sẽ lấy sai phân. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (order of intergration). Quá trình sai phân bậc d của chuỗi được thực hiện như sau:



Thông thường chuỗi sẽ dừng sau quá trình đồng tích hợp I(0) hoặc I(1). Rất ít chuỗi chúng ta phải lấy tới sai phân bậc 2.

Phương trình hồi qui ARIMA(p, d, q) có thể được biểu diễn dưới dạng:



Tổng quát thì ARIMA là mô hình kết hợp của 2 quá trình tự hồi qui và trung bình trượt. Dữ liệu trong quá khứ sẽ được sử dụng để dự báo dữ liệu trong tương lai. Trước khi huấn luyện mô hình, cần chuyển hóa chuỗi sang chuỗi dừng bằng cách lấy sai phân bậc 1 hoặc logarit.

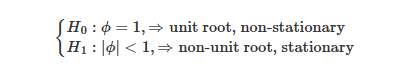
1. **Kiểm tra tính dừng của chuỗi thời gian:**

Một chuỗi thời gian có thể là dừng (stationary) hoặc không dừng (non-stationary). Một chuỗi thời gian dừng có các đặc tính thống kê không thay đổi theo thời gian. Điều này có nghĩa là các thống kê như trung bình, phương sai, tự tương quan, không thay đổi qua các dữ liệu. Hầu hết các phương pháp dự báo thống kê, bao gồm ARIMA, đều dựa trên giả định rằng chuỗi thời gian có thể được biến đổi thành chuỗi thời gian dừng thông qua một hoặc nhiều phép biến đổi. Một chuỗi thời gian dừng sẽ dễ dự đoán hơn vì có thể dự đoán rằng các đặc tính thống kê trong tương lai sẽ tương tự như trong quá khứ.

Việc dự báo dữ liệu có xu hướng có thể gặp khó khăn vì xu hướng này làm mờ đi các mẫu khác trong dữ liệu. Nếu dữ liệu có một đường xu hướng ổn định mà nó luôn quay lại theo hướng đó, nó có thể là chuỗi thời gian dừng theo xu hướng (trend-stationary), trong trường hợp này, xu hướng có thể được loại bỏ bằng cách chỉ cần vẽ một đường xu hướng và trừ đi xu hướng khỏi dữ liệu trước khi xây dựng mô hình cho nó.

Nếu dữ liệu không phải là chuỗi thời gian dừng theo xu hướng, thì nó có thể là chuỗi thời gian dừng theo sai phân (difference-stationary), và trong trường hợp này, xu hướng có thể được loại bỏ bằng cách lấy sai phân. Cách đơn giản nhất để lấy sai phân là trừ giá trị trước đó khỏi mỗi giá trị để có được một phép đo về sự thay đổi trong dữ liệu chuỗi thời gian.

Để kiếm tra phương trình đặc trưng của chuỗi có nghiệm đơn vị hay không chúng ta sử dụng kiểm định ADF. Giả thuyết null được đặt ra đó là phương trình đặc trưng có nghiệm đơn vị. Trong trường hợp p-value < 0.05 thì ta sẽ loại bỏ giả thuyết null, chấp nhận giả thuyết thay thế. Khi đó ta có thể khẳng định rằng chuỗi không có nghiệm đơn vị và có tính chất dừng:



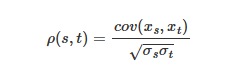
1. **Hệ số tương quan và tự tương quan riêng phần:**

**Tự tương quan (ACF - AutoCorrelation Function)**: Tự tương quan là một khái niệm quan trọng trong chuỗi thời gian. Hầu hết các chuỗi thời gian sẽ có sự tương quan với giá trị trễ của nó và các giá trị càng gần nhau thì tương quan càng mạnh hoặc các giá trị cùng thuộc 1 chu kì của chuỗi thì sẽ có tương quan cao (chẳng hạn như cùng tháng trong chu kì năm hay cùng quý trong chu kì năm).

Tính toán tự tương quan có thể giúp trả lời các câu hỏi về việc liệu dữ liệu có biểu hiện tính ngẫu nhiên hay không và mức độ liên kết giữa một quan sát và một quan sát ngay kế tiếp. Điều này có thể giúp chúng ta hiểu được mô hình nào có thể đại diện tốt nhất cho dữ liệu. Các tự tương quan thường được vẽ đồ thị để quan sát sự tương quan giữa các điểm dữ liệu, cho đến và bao gồm cả đơn vị độ trễ (lag).

Các hệ số tự tương quan tạo thành hàm tự tương quan, hay còn gọi là **ACF (Autocorrelation Function)** và thường dùng để tìm ra độ trễ của quá trình trung bình trượt MV(q) để xây dựng các mô hình như ARIMA và kiểm tra yếu tố mùa vụ.

Hệ số tự tương quan bậc s được xác định như sau:



**Tự tương quan riêng phần (PACF - Partitial AutoCorrelation Function)**: Về cơ bản tương quan riêng phần cũng là chỉ số đo lường hệ số tương quan như ACF. Tuy nhiên vẫn có sự khác biệt đó là hệ số tương quan này loại bỏ ảnh hưởng của các chuỗi độ trễ trung gian (là các chuỗi trễ xt−1,…,xt−k+1 nằm giữa xt và xt−k). Một phương trình hồi qui tuyến tính giữa chuỗi hiện tại với các chuỗi độ trễ trung gian được xây dựng nhằm đánh giá ảnh hưởng của các chuỗi độ trễ lên chuỗi hiện tại. Sau đó, để tính hệ số tương quan riêng phần chúng ta sẽ loại bỏ ảnh hưởng của các độ trễ trung gian khỏi chuỗi hiện tại bằng cách trừ đi giá trị ước lượng từ phương trình hồi qui. Lấy ví dụ: Để tính tự tương quan riêng phần PACF bậc k của chuỗi xt. Đầu tiên ta sẽ hồi qui tuyến tính xt theo các chuỗi trễ của nó là xt−1,…,xt−k. Khi đó ta thu được phương trình hồi qui tuyến tính tổng quát bậc k là:



ϵt là thành phần đại diện cho sai số. Gía trị ước lượng của mô hình đối với xt chính là:



Hệ số tự tương quan tuyến tính sau đó sẽ chính bằng:



Trong đó corr() là hàm tính hệ số tương quan.

PACF sẽ có tác dụng tìm ra hệ số bậc tự do p của quá trình tự hồi qui AR(p). Tương tự như ACF, thông qua một biểu đồ PACF về giá trị các hệ số tương quan riêng phần tương ứng với các độ trễ khác nhau, chúng ta sẽ tìm ra được các bậc tự do p phù hợp. Đó chính là vị trí mà giá trị của hệ số tương quan riêng phần nằm ngoài ngưỡng tin cậy 95% của giả thuyết hệ số tương quan riêng phần bằng 0.

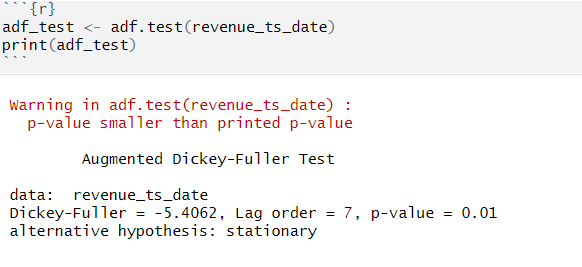
Thông thường, trong một mô hình ARIMA sẽ sử dụng hoặc là thành phần Tự Hồi Quy (AR) hoặc là thành phần Trung Bình Trượt (MA). Đồ thị ACF (Autocorrelation Function) và PACF (Partial Autocorrelation Function) thường được sử dụng để xác định thành phần nào trong số này là phù hợp nhất.

1. Ứng dụng xây dựng mô hình ARIMA dự báo doanh thu bán pizza cho 6 tháng tiếp theo

5.1. Kiểm tra tính dừng của chuỗi thời gian:





****

Kết quả kiểm định:

* Giá trị thống kê ADF (Dickey-Fuller statistic): -5.4062 - giá trị âm lớn (xa 0) thường gợi ý chuỗi có tính dừng.
* Lag order = 7 - số độ trễ được sử dụng trong kiểm định để xử lý tự tương quan.
* Giá trị p-value < 0.05, chúng ta bác bỏ giả thuyết 0 rằng chuỗi thời gian không dừng. Nghĩa là chấp nhận giả thuyết chuỗi thời gian dừng.
* Chuỗi thời gian dừng nên lấy d = 0

Kiểm tra thấy đây là chuỗi dữ liệu thời gian dừng, nên chọn mô hình ARIMA để dự đoán doanh thu bán pizza trong 6 tháng tiếp theo.

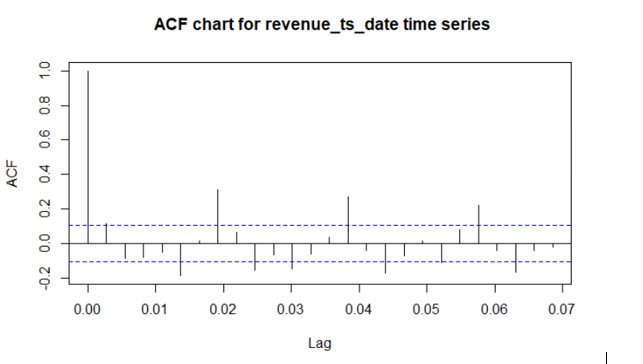
**5.2. Lựa chọn tham số ARIMA (p, d, q):**

* p: bậc của phần Tự Hồi Quy (Autoregressive - AR) của mô hình ARIMA.
* d: bậc của sự sai phân (differencing) cần thiết.
* q: bậc của phần Trung Bình Trượt (Moving Average - MA).

Thông qua đồ thị ACF và PACF, ta có thể xác định các giá trị thích hợp cho p và q.

Chúng ta có thể vẽ biểu đồ các hệ số tự tương quan ACF hàm acf của R như bên dưới:

*acf(revenue\_ts\_date, main = " ACF chart for revenue\_ts\_date time series", lag.max = 100)*



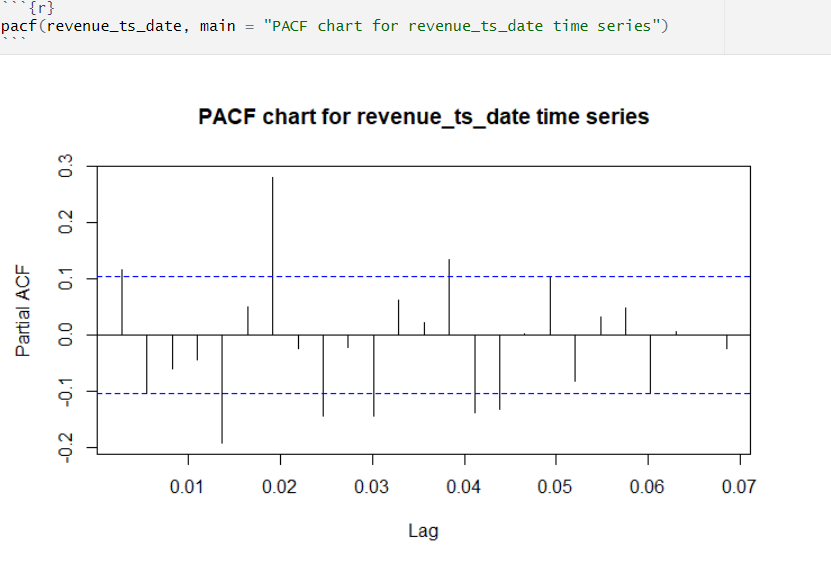
Trục hoành là độ trễ, trục tung là giá trị của hệ số tự tương quan tương ứng với độ trễ. Dãi màu xanh là khoảng tin cậy 95% để giá trị hệ số tự tương quan bằng 0. Nếu tại một độ trễ nhỏ nhất mà đoạn thẳng (vuông góc với trục hoành) mà độ dài đại diện cho giá trị của hệ số tự tương quan nằm ngoài khoảng tin cậy thì đó chính là độ trễ phù hợp lớn nhất mà ta nên lựa chọn cho quá trình trung bình trượt MV(q).

**Phân tích đồ thị:**

* **Độ trễ (lag)**:
  + **Lag = 0** chỉ là giá trị tự tương quan hoàn hảo và không liên quan đến việc xác định bậc của MA.
  + Tại **lag 2**, giá trị ACF vượt giới hạn tin cậy, cho thấy thành phần MA ảnh hưởng mạnh mẽ ở lag này.
* Lựa chọn **bậc q =2** là phù hợp.

Xác định p thông qua biểu đồ PACF:

*pacf(revenue\_ts\_date, main = "PACF chart for revenue\_ts\_date time series")*



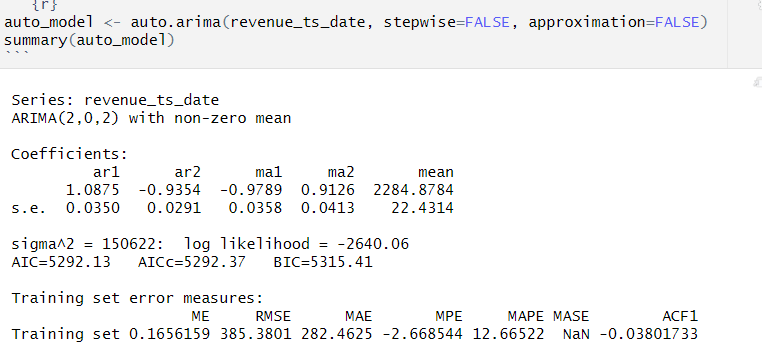
Tương tự như ACF, bậc của PACF cũng thường nhỏ hơn 5. Một giá trị PACF nằm **ngoài giới hạn này** biểu thị rằng lag đó có ảnh hưởng đáng kể và có thể được chọn làm bậc p.

Như vậy ta cũng có thể lựa chọn bậc tự do của PACF là một giá trị nào đó từ 1 đến 5.

**Dựa vào đồ thị**, thành phần AR dừng lại tại **lag 2**, vì các giá trị PACF tại lag này vượt xa giới hạn tin cậy, trong khi các lag sau không còn ảnh hưởng đáng kể.

Do đó, có thể xác định **bậc q=2, p = 2, d = 0.**

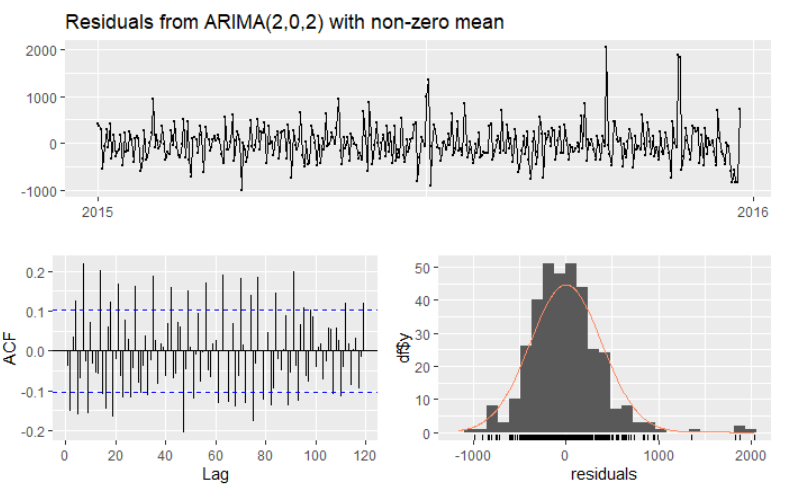
**Sử dụng cách tự động chọn tham số ARIMA (p, d, q) trong R:**

’

Tham số được chọn là ARIMA (2,0,2). Giá trị của AIC càng nhỏ thì mô hình càng phù hợp. ACF1 gần 0 cho thấy phần dư không có tự tương quan đáng kể, điều này là tốt vì nó cho thấy mô hình đã nắm bắt được cấu trúc của chuỗi thời gian.

**5.3. Kiểm tra sự phù hợp của mô hình ARIMA**

Có thể kiểm tra các giá trị residuals (phần dư) để đảm bảo rằng mô hình không có vấn đề với giả thuyết chuỗi thời gian.

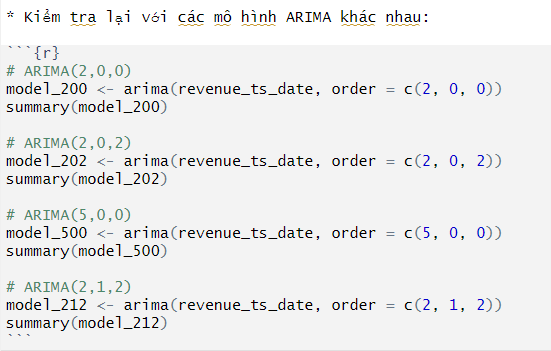


* Kết quả kiểm định: p-value < 0.05: Chúng ta bác bỏ giả thuyết gốc, nghĩa là phần dư vẫn có tự tương quan.
* Điều này gợi ý rằng mô hình ARIMA(2,0,2) chưa đủ để mô hình hóa hoàn toàn dữ liệu, và có thể cần điều chỉnh hoặc thử nghiệm thêm các cấu trúc mô hình khác.
* Phân phối phần dư dường như hơi lệch với đuôi dài, đặc biệt có một số phần dư cực lớn. Điều này cho thấy phân phối của phần dư không hoàn toàn tuân theo chuẩn. Các giá trị tự tương quan của phần dư không nằm hoàn toàn trong giới hạn tin cậy (đường chấm xanh), đặc biệt ở một số độ trễ (lags).

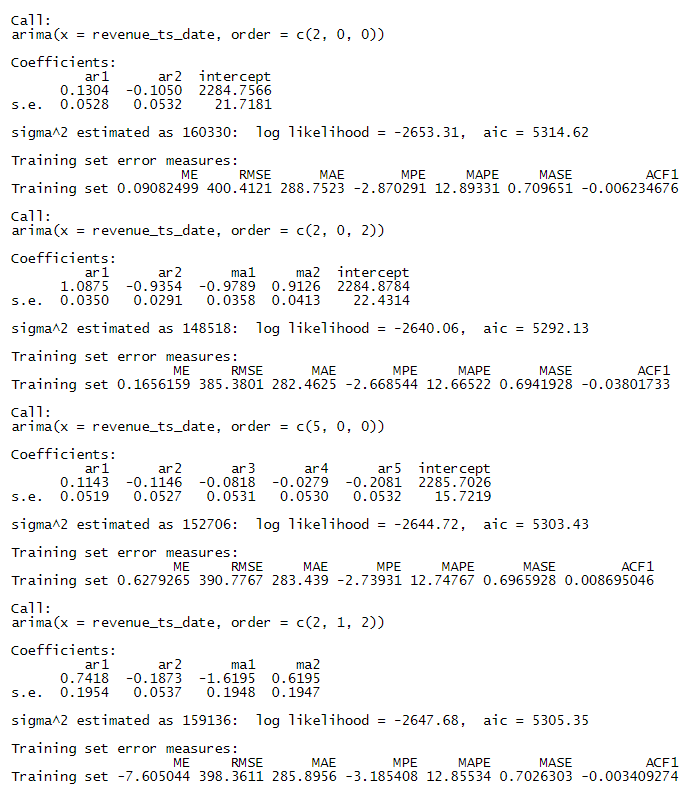
**5.4. So sánh với các mô hình ARIMA khác**

Kết hợp giữa bậc của p và q và giá trị của d = 0 do chuỗi đã là một chuỗi dừng ta có thể thu được một số kịch bản để so sánh với mô hình ARIMA (2,0,2):

* ARIMA(2, 0, 2)
* ARIMA(2, 0, 0)
* ARIMA(5, 0, 0)
* ARIMA(2, 1, 2)



Kết quả:



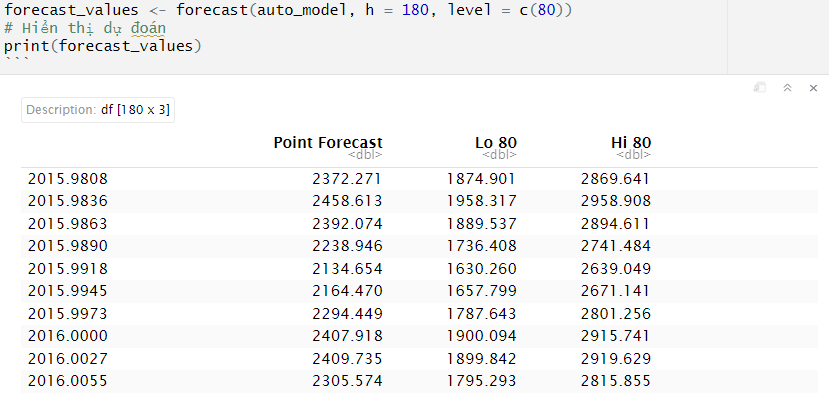
Có ba chỉ số quan trọng để đánh giá độ chính xác của mô hình ARIMA:

1. **Chỉ số AIC - Akaike Information Criteria:** Đây là chỉ số được sử dụng rộng rãi để chọn các yếu tố dự báo trong các mô hình hồi quy và cũng hữu ích cho việc xác định bậc của mô hình ARIMA. Điểm AIC càng thấp thì mô hình càng tốt, vì vậy chúng ta sẽ ưu tiên mô hình có điểm AIC thấp hơn.
2. **Chỉ số BIC - Bayesian Information Criterion**: Đây là một chỉ số khác để lựa chọn mô hình, với việc phạt sự phức tạp nhiều hơn so với AIC. Giống như AIC, các mô hình có điểm BIC thấp thường được ưa chuộng hơn các mô hình có điểm BIC cao. Nếu mô hình sử dụng cho dự báo dài hạn, thì BIC có thể là lựa chọn ưu tiên, trong khi đối với dự báo ngắn hạn, AIC có thể là lựa chọn tốt hơn.
3. **Giá trị sigma bình phương (sigma²)**: Đây là phương sai của các phần dư mô hình. Hệ số sigma mô tả sự biến động của quá trình giả thuyết. Nếu dữ liệu của có sự biến động lớn nhưng giá trị sigma bình phương rất thấp, hoặc ngược lại, dữ liệu ít biến động nhưng giá trị sigma bình phương cao, điều này là dấu hiệu cho thấy mô hình không nắm bắt được quá trình sinh ra dữ liệu thực tế một cách chính xác.

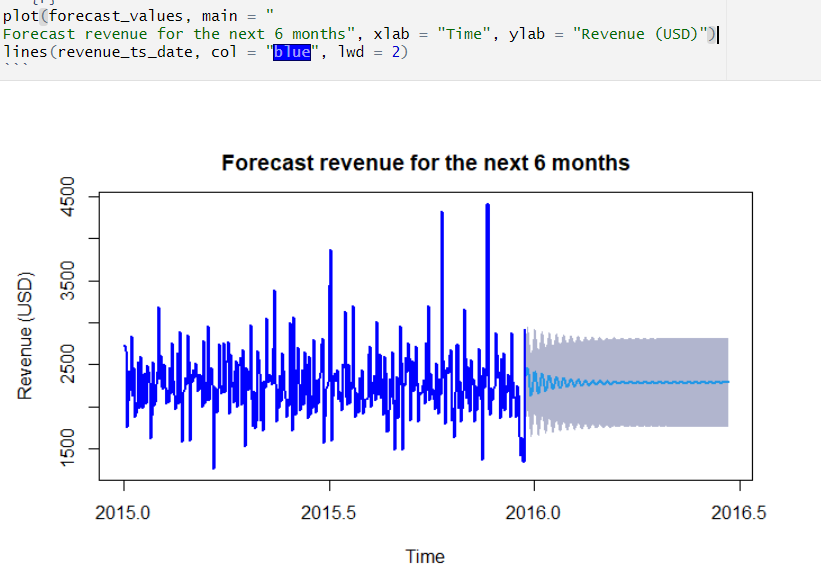
* So sánh các chỉ số AIC cho thấy mô hình ARIMA (2,0,2) vẫn tốt hơn.

**5.5. Xây dựng mô hình ARIMA dự đoán doanh thu 6 tháng tiếp theo:**

Dự đoán doanh thu trong 6 tháng (180 ngày) tiếp theo. Do điểm dữ liệu doanh thu theo tháng quá ít để xây dựng mô hình, nên sẽ dự đoán theo ngày, với số ngày là 180 ngày.



Trực quan hóa số liệu dự đoán:



**\*\*Nhận xét:\*\***

1. Xu hướng dự báo: Doanh thu pizza trong 6 tháng dự báo có xu hướng ổn định, không tăng hoặc giảm mạnh theo thời gian. Điều này thể hiện qua đường trung bình (màu xanh), dao động quanh một mức giá trị cố định, không có sự thay đổi đáng kể về xu hướng.

2. Biến động của dữ liệu: Trong dữ liệu lịch sử, doanh thu pizza có những giai đoạn dao động lớn và xuất hiện các đỉnh cao đột ngột. Tuy nhiên, dự báo không phản ánh các biến động lớn đó, mà chỉ nhắm đến mức trung bình chung, làm giảm khả năng dự đoán các đột biến doanh thu (nếu có).

3. Về độ tin cậy: Với khoảng tin cậy khá rộng, đặc biệt ở cuối giai đoạn dự báo, điều này có thể ám chỉ rằng mô hình gặp khó khăn trong việc nắm bắt toàn bộ đặc điểm của dữ liệu, đặc biệt là những biến động bất thường trong lịch sử.

Do đó, dự đoán này nên được sử dụng thận trọng, đặc biệt trong việc ra quyết định dài hạn.

4. Ổn định: Nếu xu hướng trung bình tiếp tục, doanh nghiệp có thể dự đoán doanh thu tương đối ổn định mà không cần phải đối mặt với biến động lớn trong ngắn hạn nhưng cũng có rủi ro tiềm ẩn: Các sự kiện bất thường (như khuyến mãi, nhu cầu mùa cao điểm) có thể không được phản ánh đầy đủ trong mô hình. Điều này cần được cân nhắc khi lập kế hoạch kinh doanh.

**Đánh giá về hiệu quả việc dùng mô hình ARIMA cho phân tích doanh thu trong 6 tháng tiếp theo:**

**1. Hiệu quả dự đoán từ đồ thị:**

Đường dự báo (forecast line): Phần dự báo có xu hướng ổn định và không có sự biến động lớn trong tương lai, thể hiện mô hình không dự đoán được các dao động hoặc bất thường tiềm tàng trong doanh thu.

Dải tin cậy (confidence intervals): Dải màu xám xung quanh dự báo biểu thị khoảng tin cậy. Nó mở rộng dần khi dự đoán xa hơn, phản ánh sự không chắc chắn tăng dần. Tuy nhiên, dải tin cậy khá lớn, cho thấy mô hình chưa thực sự tốt trong việc nắm bắt các dao động phức tạp của chuỗi thời gian doanh thu.

**2. Phân tích phần dư (Residuals):**

**Ljung-Box Test:**

Kết quả kiểm định: 𝑝 – value < 2.2𝑒−16, cho thấy phần dư không độc lập. Điều này chỉ ra rằng mô hình ARIMA chưa xử lý hoàn toàn các tự tương quan trong chuỗi thời gian.

**Phần dư**: Từ các đồ thị ACF và PACF của phần dư, có vẻ vẫn còn sự tự tương quan ở các lag, đặc biệt ở những lag nhỏ (lag 1, lag 2,...).

Điều này chỉ ra rằng mô hình chưa tối ưu trong việc mô hình hóa toàn bộ cấu trúc chuỗi thời gian, có thể do thiếu các biến số ngoại sinh hoặc yếu tố mùa vụ chưa được xử lý tốt.

**3. Đánh giá thông số mô hình:**

AR(2), MA(2) có ý nghĩa thống kê, nhưng sigma^2 (150622) khá lớn, biểu thị sai số dự đoán còn cao.

BIC và AIC đều cho thấy mức độ phức tạp tương đối của mô hình (AIC = 5292.13). Tuy nhiên, việc tối ưu thêm mô hình có thể giảm giá trị này.

**4. Tổng quan về hiệu quả:**

**Ưu điểm:**

Mô hình ARIMA có thể nắm bắt được xu hướng tổng quan và phần nào dự đoán được các mức giá trị trung bình của doanh thu trong tương lai.

Phù hợp để dự báo ngắn hạn với dữ liệu không quá phức tạp hoặc ít dao động mạnh.

**Nhược điểm:**

Chưa xử lý tốt tính mùa vụ hoặc các yếu tố ngoại sinh có ảnh hưởng lớn đến doanh thu.

Độ tin cậy của dự báo trong dài hạn thấp, thể hiện qua dải tin cậy lớn.

Phần dư không hoàn toàn ngẫu nhiên, cho thấy cần xem xét các mô hình khác (ví dụ: SARIMA hoặc sử dụng các biến bổ sung).

**5. Kết luận và đề xuất tối ưu:**

**Cải thiện mô hình:**

Đánh giá thêm dữ liệu:

Tăng số lượng dữ liệu lịch sử hoặc kiểm tra các đặc điểm phức tạp khác của chuỗi (dao động không theo quy luật).

**Hiệu quả:**

ARIMA phù hợp cho dự báo ngắn hạn, nhưng nếu muốn độ chính xác cao hơn cho các dự báo 6 tháng hoặc dài hạn, cần tối ưu thêm mô hình hoặc sử dụng các phương pháp khác (như mô hình machine learning hoặc SARIMA).

# KẾT LUẬN

Mục tiêu của đề tài là thực hiện bài toán phân tích dữ liệu và xây dựng mô hình dự đoán với ngôn ngữ R trên tập dữ liệu bán hàng của một cửa hàng pizza. Đồ án của nhóm đã đạt được mục tiêu đề ra bằng cách: Hiểu dữ liệu và bài toán; Tiền xử lý dữ liệu như kiểm tra giá trị thiếu, chuyển sang kiểu dữ liệu phù hợp, tạo các biến mới cần thiết…; Cung cấp các đánh giá và trực quan hóa kết quả phân tích nhu cầu tiêu thụ pizza, nguyên liệu, doanh số, doanh thu bằng nhiều loại biểu đồ đa dạng; Xây dựng được mô hình dự đoán doanh số với độ chính xác cao và sai số nhỏ; Dự đoán doanh thu trong 6 tháng tiếp theo bằng mô hình ARIMA. Mặc dù đồ án đã sử dụng một tập dữ liệu khá lớn, tuy nhiên thông tin dữ liệu là từ các đơn đặt hàng pizza trong năm 2015, mang tính chất thực hành phân tích. Ngoài ra, tập dữ liệu chứa ít thông tin về khách hàng, gây khó khăn trong việc thực hiện các phân tích liên quan như phân khúc khách hàng, phân tích hành vi mua sắm. Qua những hạn chế đã đề cập, để phát triển đề tài này nhóm có thể thu thập nhiều thông tin liên quan đến khách hàng, sử dụng dữ liệu mới nhất từ các hệ thống bán hàng thời gian thực nhằm nâng cao giá trị của các phân tích.

# TÀI LIỆU THAM KHẢO

1. Winton Clem (2024), Analysis of Data and the Field of Data Science, First edition
2. Peter J. Brockwell & Richard A. Davis (2002), Introduction to Time Series and Forecasting, 2nd ed, Nxb. Springer.
3. Joshua Noble, What are ARIMA models?, ibm.com, <https://www.ibm.com/topics/arima-model>, truy cập ngày 10/11/2024.
4. G.E. Box, G.M. Jenkins, G.C. Reinsel and G.M. Ljung (2016), Time series analysis: forecasting and control, 4th ed, John Wiley & Sons.
5. Phạm Đình Khánh (2019), Mô hình ARIMA trong time series, <https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html#42-ch%E1%BB%89-s%E1%BB%91-aic---akaike-information-criteria>, truy cập ngày 12/11/2024.
6. Prabhakaran Selva (2019), ARIMA Model – Complete Guide to Time Series Forecasting in Python, Machine Learning Plus, machinelearningplus.com, <https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/>, truy cập ngày 15/11/2024.
7. Deepanshu Bhalla, Random Forest in R: A Step-by-Step Guide, Listen Data, <https://www.listendata.com/2014/11/random-forest-with-r.html>, truy cập ngày 10/11/2024
8. Rebecca Bevans (2020), ANOVA in R, Scribbr, <https://www.scribbr.com/statistics/anova-in-r/>, truy cập ngày 04/11/2024. 9.Zach Bobbitt (2022), How to Perform a Kruskal-Wallis Test in R, Statology, <https://www.statology.org/kruskal-wallis-test-in-r/>, truy cập ngày 10/11/2024.
9. Bùi Tiến Trung – Ngày 28 tháng 6 năm 2021, <https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0>
10. XGBoost Documentation, <https://xgboost.readthedocs.io/en/latest/R-package/xgboostPresentation.html>
11. GreekForGreeks, <https://www.geeksforgeeks.org/gradient-boosting-in-r/>
12. TomorrowMarketers, Biểu đồ boxplot là gì và đọc hiểu biểu đồ này như thế nào?, <https://blog.tomorrowmarketers.org/bieu-do-boxplot/>, truy cập ngày 25/11/2024
13. TomorrowMarketers, <https://blog.tomorrowmarketers.org/time-series-analysis/>, Dữ liệu chuỗi thời gian (Time Series) là gì? Phân tích dữ liệu chuỗi thời gian như thế nào?, truy cập ngày 23/11/2024

# BẢNG PHÂN CÔNG VÀ ĐÁNH GIÁ

## Bảng phân công nhiệm vụ



## Thành viên nhận xét và đánh giá

