| | **BỘ CÔNG THƯƠNG**  **TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP** | **KHOA: KHOA HỌC ỨNG DỤNG** | | --- | --- |   **BÁO CÁO ĐỒ ÁN 1**  **“SỬ DỤNG NGÔN NGỮ R ĐỂ THĂM DÒ VÀ TRỰC QUAN DỮ LIỆU VỀ THÀNH PHẦN DINH DƯỠNG CÓ TRONG ĐỒ ĂN CỦA HÃNG MCDONALD'S”**  Sinh viên thực hiện: Lương Đức Thắng  Nguyễn Đăng Thanh  Lê Thị Linh Phương  Phạm Thị Ngọc Tú  Trần Thị Thu Trang  Giảng viên hướng dẫn: Ths. Trần Chí Lê    **HÀ NỘI - 2024** |
| --- | --- | --- |

| **BỘ CÔNG THƯƠNG**  **TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP** | **KHOA: KHOA HỌC ỨNG DỤNG** |
| --- | --- |

**BÁO CÁO ĐỒ ÁN 1**

**“SỬ DỤNG NGÔN NGỮ R ĐỂ THĂM DÒ VÀ TRỰC QUAN DỮ LIỆU VỀ THÀNH PHẦN DINH DƯỠNG CÓ TRONG ĐỒ ĂN CỦA HÃNG MCDONALD'S”**

Sinh viên thực hiện: Lương Đức Thắng

Nguyễn Đăng Thanh

Lê Thị Linh Phương

Phạm Thị Ngọc Tú

Trần Thị Thu Trang

Giảng viên hướng dẫn: Ths. Trần Chí Lê

**HÀ NỘI - 2024**

# **LỜI CẢM ƠN**

Trước tiên, nhóm chúng em xin bày tỏ lòng biết ơn sâu sắc đến quý Thầy Cô trong khoa Khoa học Ứng dụng của Trường Đại học Kinh tế - Kỹ thuật Công nghiệp đã tận tình giảng dạy và truyền đạt kiến thức quý báu trong suốt quá trình học tập và nghiên cứu tại trường.

Đặc biệt, em xin gửi lời cảm ơn chân thành đến Thầy Trần Chí Lê, người đã trực tiếp hướng dẫn và hỗ trợ nhóm chúng em trong suốt quá trình thực hiện đồ án này. Sự chỉ bảo tận tình, những góp ý quý báu và sự kiên nhẫn của Thầy đã giúp chúng em hoàn thành đồ án này.

Chúng em cũng xin cảm ơn đến các bạn trong lớp đã luôn chia sẻ, động viên và hỗ trợ nhóm trong suốt thời gian thực hiện đồ án. Những ý kiến đóng góp và sự giúp đỡ của các bạn đã giúp nhóm em có thêm động lực và cải thiện chất lượng công việc của mình.

Cuối cùng, nhóm chúng em xin gửi lời cảm ơn đến gia đình và bạn bè đã luôn ủng hộ, động viên và tạo điều kiện tốt nhất cho chúng em trong suốt thời gian học tập và hoàn thành đồ án này.

Mặc dù đã cố gắng hết sức, nhưng do kiến thức và kinh nghiệm còn hạn chế, báo cáo đồ án này không thể tránh khỏi những thiếu sót. Chúng em kính mong nhận được sự góp ý từ quý Thầy Cô và các bạn để báo cáo của em được hoàn thiện hơn.

Nhóm chúng em xin chân thành cảm ơn!

## 

# **MỤC LỤC**

[**LỜI CẢM ƠN 3**](#_gjdgxs)

[**MỤC LỤC 4**](#_1fob9te)

[**DANH MỤC HÌNH ẢNH 6**](#)

[**DANH MỤC BẢNG 9**](#_2et92p0)

[**LỜI MỞ ĐẦU 10**](#_4d34og8)

[**I. LÝ DO CHỌN ĐỀ TÀI 10**](#_2s8eyo1)

[**II. PHƯƠNG PHÁP THỰC HIỆN 12**](#_17dp8vu)

[**CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ THƯ VIỆN GGPLOT VÀ PHÂN TÍCH THỐNG KÊ 13**](#_3rdcrjn)

[**1.1. THĂM DÒ DỮ LIỆU BẰNG BIỂU ĐỒ GGPLOT2 13**](#_26in1rg)

[**1.2. TÓM TẮT KẾT QUẢ THEO SUY DIỄN THỐNG KÊ 21**](#_z337ya)

[**1.2.2. Thống kê suy diễn trong các bài toán kiểm định 22**](#_1y810tw)

[**1.2.3. Thống kê suy diễn trong các bài toán phân tích tương quan 27**](#_1pxezwc)

[**CHƯƠNG 2: PHÂN TÍCH THĂM DÒ DỮ LIỆU 31**](#_1hmsyys)

[**2.1. GIỚI THIỆU BÀI TOÁN 31**](#_41mghml)

[**2.1.1. Giới thiệu bài toán 31**](#_2grqrue)

[**2.1.2. Thu thập dữ liệu 31**](#_vx1227)

[**2.1.3. Phạm vi và giới hạn 32**](#_3fwokq0)

[**2.1.4. Phương pháp và công cụ 32**](#_1v1yuxt)

[**2.1.5. Giải thích các biến dữ liệu 32**](#_4f1mdlm)

[**2.2. XỬ LÝ DỮ LIỆU 33**](#_2u6wntf)

[**2.2.1. Thu thập dữ liệu 33**](#_19c6y18)

[**2.2.2. Kiểm tra dữ liệu 33**](#_28h4qwu)

[**2.2.3. Làm sạch dữ liệu 35**](#_2lwamvv)

[**2.3.THỐNG KÊ MÔ TẢ 36**](#_3l18frh)

[**2.3.1. Tính toán các thống kê cơ bản 36**](#_206ipza)

[**2.3.2. Công thức chung để tính toán các biến và vẽ biểu đồ 39**](#_1egqt2p)

[**2.4. PHÂN TÍCH MỐI QUAN HỆ GIỮA CÁC BIẾN 46**](#_3cqmetx)

[**2.4.1. Tính ma trận tương quan 46**](#_1rvwp1q)

[**2.4.2. Trực quan hóa mối liên hệ giữa các biến 48**](#_2r0uhxc)

[**CHƯƠNG 3 : TRỰC QUAN HÓA DỮ LIỆU 58**](#_xvir7l)

[**3.1. GIỚI THIỆU VỀ TRỰC QUAN HÓA DỮ LIỆU 58**](#_3hv69ve)

[**3.1.1.Khái niệm 58**](#_1x0gk37)

[**3.1.2. Mục đích 58**](#_4h042r0)

[**3.1.3. Phân loại 58**](#_2w5ecyt)

[**3.2. BIỂU ĐỒ VỀ TRỰC QUAN HÓA DỮ LIỆU 58**](#_1baon6m)

[**3.2.1. Biểu đồ tròn phân phối mục theo danh sách thực đơn 58**](#_3vac5uf)

[**3.2.2. Biểu đồ Bar chart 60**](#_pkwqa1)

[**3.2.3. Biểu đồ Scatter plot 69**](#_40ew0vw)

[**3.2.4. Biểu đồ Boxplot 77**](#_meukdy)

[**3.3. PHÂN TÍCH HỒI QUY TUYẾN TÍNH 86**](#_3x8tuzt)

[**3.3.1. Hệ số tương quan 86**](#_2ce457m)

[**3.3.2. Biểu đồ hồi quy 88**](#_3bj1y38)

[**3.4. PHÂN TÍCH THÀNH PHẦN CHÍNH PCA 98**](#_wnyagw)

[**3.4.1. Chuẩn hóa dữ liệu 98**](#_3gnlt4p)

[**3.4.2. Trực quan hóa và giải thích 99**](#_4fsjm0b)

[**3.4.3. Đồ thị của các biến 101**](#_3u2rp3q)

[**CHƯƠNG 4: TỔNG KẾT 105**](#_11si5id)

[**TÀI LIỆU THAM KHẢO 107**](#_3ls5o66)

# 

# **DANH MỤC HÌNH ẢNH**

[*Hình 1.1: Minh họa quy trình thêm các lớp hàm phân tích trong ggplot 1*](#_lnxbz9)*3*

[*Hình 1.2: Cách gán biến số vào các trục 1*](#_1ksv4uv)*4*

[*Hình 1.3: Xuất bản biểu đồ qua lệnh Import trong Rstudio 2*](#_2jxsxqh)*1*

[*Hình 1.4: Cú pháp kiểm định cho kỳ vọng một mẫu 2*](#_2xcytpi)*4*

[*Hình 1.5: Cú pháp kiểm định giả thuyết cho tỷ lệ một mẫu 2*](#_1ci93xb)*5*

[*Hình 1.6: Kết quả của trung bình hai mẫu 2*](#_3whwml4)*5*

[*Hình 1.7: Kết quả của so sánh phương sai 2*](#_qsh70q)*6*

[*Hình 1.8: Thủ tục kiểm định 2*](#_3as4poj)*7*

[*Hình 1.9: Đọc file marketing.csv 2*](#_49x2ik5)*8*

[*Hình 1.10: Cú pháp tính hệ số tương quan Pearson trong R 2*](#_147n2zr)*9*

[*Hình 1.11: Cú pháp tính hệ số tương quan Spearman trong R 2*](#_3o7alnk)*9*

[*Hình 1.12: Cú pháp tính hệ số tương quan Kendall trong R*](#_23ckvvd) *30*

[*Hình 1.13: Cú pháp tính kiểm định trong R*](#_32hioqz) *30*

[*Hình 2.1: Biểu đồ về số lượng món ăn theo Menu 3*](#_2zbgiuw)*8*

[*Hình 2.2: Biểu đồ thống kê trung bình lượng Total fat theo Menu*](#_3ygebqi) *40*

[*Hình 2.3: Biểu đồ phân phối của Total carbohydrate 4*](#_2dlolyb)*2*

[*Hình 2.4: Biểu đồ phân phối ECDF của biến Energy (kCal) 4*](#_sqyw64)*4*

[*Hình 2.5: Ma trận tương quan giữa các thành phần dinh dưỡng 4*](#_4bvk7pj)*7*

[*Hình 2.6: Ma trận tương quan thông qua biểu đồ tròn và màu sắc*](#_1664s55) *50*

[*Hình 2.7: Ma trận tương quan thông qua số liệu và màu sắc*](#_3q5sasy) *51*

[*Hình 2.8: Ma trận tương quan thể hiện bởi màu sắc 5*](#_25b2l0r)*2*

[*Hình 2.9: Ma trận tương quan biểu diễn thông qua biểu đồ bánh 5*](#_kgcv8k)*3*

[*Hình 2.10: Biểu đồ Scatter plot giữa Energy (kCal) và Total fat (g) 5*](#_1jlao46)*5*

[*Hình 2.11: Biểu đồ thể hiện mối liên hệ giữa Sodium và Added Sugars 5*](#_2iq8gzs)*7*

[*Hình 3.1: Biểu đồ phân phối mục theo danh sách thực đơn*](#_2afmg28) *59*

[*Hình 3.2: Biểu đồ bar chart cho energy.*](#_39kk8xu) *60*

[*Hình 3.3: Biểu đồ bar chart cho protein 6*](#_1opuj5n)*1*

*Hình 3.4: Biểu đồ bar chart cho total fat**62*

[*Hình 3.5: Biểu đồ bar chart cho sat fat… 6*](#_2nusc19)*3*

[*Hình 3.6: Biểu đồ bar chart cho Trans Fat 6*](#_1302m92)*4*

[*Hình 3.7: Biểu đồ bar chart cho cholesterols 6*](#_3mzq4wv)*5*

[*Hình 3.8: Biểu đồ bar chart cho total carbohydrate 6*](#_2250f4o)*6*

[*Hình 3.9: Biểu đồ bar chart cho total sugar 6*](#_haapch)*7*

[*Hình 3.10: Biểu đồ bar chart cho added sugar 6*](#_319y80a)*8*

[*Hình 3.11: Biểu đồ bar chart cho sodium*](#_1gf8i83) *69*

[*Hình 3.12: Biểu đồ scatter giữa energy và protein*](#_2fk6b3p) *70*

[*Hình 3.13: Biểu đồ scatter giữa energy và total fat*](#_upglbi) *71*

[*Hình 3.14: Biểu đồ scatter giữa energy và sat fat 7*](#_3ep43zb)*2*

[*Hình 3.15: Biểu đồ scatter giữa energy và trans fat 7*](#_1tuee74)*3*

[*Hình 3.16: Biểu đồ scatter giữa energy và cholesterols 7*](#_4du1wux)*3*

[*Hình 3.17: Biểu đồ scatter giữa energy và total carbohydrate 7*](#_2szc72q)*4*

[*Hình 3.18: Biểu đồ scatter giữa energy và total sugars 7*](#_184mhaj)*5*

[*Hình 3.19: Biểu đồ scatter giữa energy và added sugars 7*](#_3s49zyc)*6*

[*Hình 3.20: Biểu đồ scatter giữa energy và sodium 7*](#_279ka65)*7*

[*Hình 3.21: Biểu đồ boxplot của energy 7*](#_36ei31r)*8*

[*Hình 3.22: Biểu đồ boxplot của protein*](#_1ljsd9k) *79*

[*Hình 3.23: Biểu đồ boxplot của total fat*](#_45jfvxd) *79*

[*Hình 3.24: Biểu đồ boxplot của sat fat*](#_2koq656) *80*

[*Hình 3.25: Biểu đồ boxplot của trans fat*](#_zu0gcz) *81*

[*Hình 3.26: Biểu đồ boxplot của cholesterols*](#_3jtnz0s) *82*

[*Hình 3.27: Biểu đồ boxplot của total carbohydrate 8*](#_1yyy98l)*3*

[*Hình 3.28: Biểu đồ boxplot của total sugars 8*](#_4iylrwe)*4*

[*Hình 3.29: Biểu đồ boxplot của added sugars 8*](#_2y3w247)*4*

[*Hình 3.30: .Biểu đồ boxplot của sodium 8*](#_1d96cc0)*5*

[*Hình 3.31: Biểu đồ hồi quy tuyến tính giữa biến Energy và Protein*](#_1qoc8b1) *89*

[*Hình 3.32: Biểu đồ hồi quy tuyến tính giữa Energy và Total fat*](#_4anzqyu) *90*

[*Hình 3.33:Biểu đồ hồi quy tuyến tính giữa Energy và Sat fat*](#_2pta16n) *91*

[*Hình 3.34: Biểu đồ hồi quy tuyến tính giữa Energy và Total carbohydrate*](#_14ykbeg) *92*

[*Hình 3.35: Biểu đồ hồi quy tuyến tính giữa Energy và Sodium 9*](#_3oy7u29)*3*

[*Hình 3.36: Biểu đồ hồi quy tuyến tính giữa Protein và Total fat 9*](#_243i4a2)*3*

[*Hình 3.37: Biểu đồ hồi quy tuyến tính giữa Protein và Sodium 9*](#_j8sehv)*4*

[*Hình 3.38: Biểu đồ hồi quy tuyến tính giữa Total fat và Sat fat 9*](#_338fx5o)*5*

[*Hình 3.39: Biểu đồ hồi quy tuyến tính giữa Total fat và Sodium 9*](#_1idq7dh)*6*

[*Hình 3.40: Biểu đồ hồi quy tuyến tính giữa Total carbohydrate và Total fat 9*](#_42ddq1a)*7*

[*Hình 3.41: Biểu đồ hồi quy tuyến tính giữa Total Sugars và Added Sugars 9*](#_2hio093)*7*

[*Hình 3.42: Biểu đồ scree plot thể hiện giá trị riêng*](#_1a346fx) *101*

[*Hình 3.43: Vòng tròn tương quan của các biến*](#_odc9jc) *102*

[*Hình 3.44: Biểu đồ thể hiện chất lượng thể hiện của 2 thành phần chính 10*](#_47hxl2r)*3*

[*Hình 3.45 : Biểu đồ vòng tròn tương quan thể hiện màu sắc theo giá trị 10*](#_2mn7vak)*4*

# **DANH MỤC BẢNG**

[*Bảng 1.1: Dữ liệu quan sát về các mẫu ô tô (nguồn: ggplot2) 1*](#_35nkun2)*4*

[*Bảng 1.2: Các đối số hay sử dụng cho việc căn chỉnh trong hàm theme. 1*](#_44sinio)*7*

[*Bảng 1.3: Các hàm tính thống kê mô tả cơ bản trong R 2*](#_3j2qqm3)*2*

[*Bảng 1.4: Dữ liệu quan sát số lượt quảng cáo, (nguồn: internet) 2*](#_2p2csry)*8*

[*Bảng 2.1: Dữ liệu dinh dưỡng từ thực đơn McDonald’s 3*](#_3tbugp1)*3*

[*Bảng 2.2: Hiển thị 6 dòng đầu của data 3*](#_nmf14n)*4*

[*Bảng 2.3: Hiển thị 6 dòng cuối của data 3*](#_37m2jsg)*4*

[*Bảng 2.4: Hiển thị cấu trúc của data 3*](#_1mrcu09)*4*

[*Bảng 2.5: Tóm tắt thống kê cho từng cột trong data 3*](#_46r0co2)*5*

[*Bảng 2.6: Kiểm tra số lượng NA của các biến 3*](#_111kx3o)*5*

[*Bảng 2.7: Tóm tắt thống kê mô tả các nhóm 3*](#_4k668n3)*7*

[*Bảng 2.8: Các biến có giá trị tương quan cao nhất 5*](#_34g0dwd)*4*

[*Bảng 2.9: Các biến có giá trị tương quan thấp nhất 5*](#_43ky6rz)*6*

[*Bảng 3.1: Kết quả hệ số tương quan giữa các biến 8*](#_rjefff)*7*

[*Bảng 3.2: Kết quả phân tích thành phần chính (PCA)*](#_1vsw3ci) *99*

[*Bảng 3.3: Giá trị riêng*](#_2uxtw84) *100*

[*Bảng 3.4: Kết quả phân tích thành phần chính cho các biến*](#_2981zbj) *101*

[*Bảng 3.5: 4 dòng của var$cos2*](#_38czs75) *103*

## 

# **LỜI MỞ ĐẦU**

## **LÝ DO CHỌN ĐỀ TÀI**

McDonald's là một trong những chuỗi thực phẩm hàng đầu trên thế giới với lượng người theo dõi đông đảo. Với hơn 38.000 cửa hàng trên khắp thế giới và hàng triệu khách hàng mỗi ngày, McDonald's đã trở thành một biểu tượng của văn hóa ẩm thực toàn cầu. Tuy nhiên, với sự phát triển nhanh chóng của công nghệ và sự quan tâm ngày càng tăng về dinh dưỡng và sức khỏe, việc tiến hành phân tích dữ liệu khám phá và tìm hiểu về phân tích dinh dưỡng trong sản phẩm của McDonald's trở nên cần thiết hơn bao giờ hết.

Phân tích dữ liệu khám phá là quá trình khám phá và phân tích dữ liệu để hiểu rõ về chúng mà không cần áp dụng mô hình dự đoán cụ thể. Trong trường hợp của McDonald's, việc tiến hành EDA có thể giúp chúng ta hiểu rõ hơn về hành vi tiêu dùng của khách hàng, xu hướng ưa thích của họ và cách mà McDonald's có thể cải thiện dịch vụ của mình. Bằng cách phân tích dữ liệu từ số lượng lớn các giao dịch mua hàng, chúng ta có thể nhận ra những mô hình tiêu dùng, thói quen ăn uống và nhu cầu của khách hàng, từ đó đưa ra các chiến lược kinh doanh hiệu quả hơn.

Ngoài ra, việc tìm hiểu về phân tích dinh dưỡng trong sản phẩm của McDonald's cũng rất quan trọng. Với một menu đa dạng từ hamburger, khoai tây chiên đến salad và nước uống, việc đánh giá giá trị dinh dưỡng của từng sản phẩm giúp khách hàng có thể chọn lựa một cách thông minh và lành mạnh hơn. Bằng cách phân tích dinh dưỡng, chúng ta có thể biết được lượng calo, chất béo, protein và carbohydrate trong mỗi sản phẩm, từ đó giúp người tiêu dùng có thể lựa chọn sản phẩm phù hợp với nhu cầu dinh dưỡng của mình.

Thêm vào đó, việc phân tích dinh dưỡng cũng giúp McDonald's cải thiện chất lượng sản phẩm và đáp ứng nhu cầu ngày càng tăng về sức khỏe của khách hàng. Bằng cách tăng cường dinh dưỡng trong các sản phẩm, McDonald's có thể thu hút thêm đông đảo khách hàng quan tâm đến sức khỏe và dinh dưỡng, đồng thời cũng nâng cao hình ảnh và uy tín của mình trên thị trường.

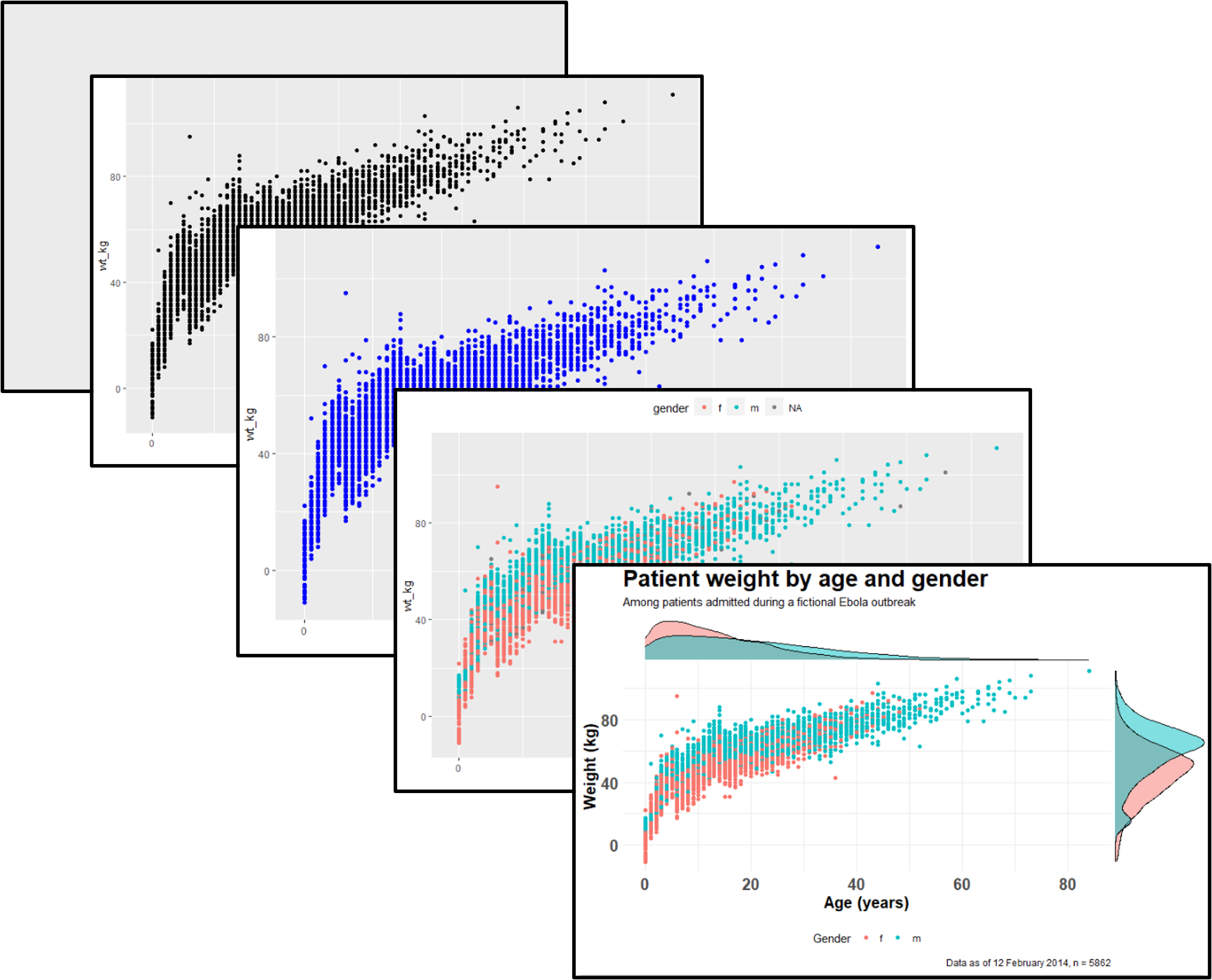
Do đó, việc tiến hành phân tích dữ liệu khám phá và tìm hiểu về phân tích dinh dưỡng trong sản phẩm của McDonald's là rất cần thiết để cải thiện chất lượng dịch vụ, đáp ứng nhu cầu của người tiêu dùng và nâng cao uy tín của thương hiệu. Việc này không chỉ mang lại lợi ích kinh doanh mà còn đóng góp vào việc thúc đẩy sự phát triển bền vững của ngành công nghiệp thực phẩm và dinh dưỡng trên toàn cầu.

## **PHƯƠNG PHÁP THỰC HIỆN**

1. Xác định yêu cầu của đồ án: Đầu tiên, chúng tôi xác định rõ yêu cầu của đồ án, bao gồm: mục tiêu, phạm vi và các chức năng cần thực hiện.
2. Thu thập dữ liệu: Sau khi đã xác định yêu cầu, chúng tôi thu thập dữ liệu cần thiết cho đồ án. Dữ liệu có thể được thu thập từ các nguồn khác nhau như cơ sở dữ liệu, tệp tin CSV, API, vv,...
3. Tiến hành xử lý dữ liệu: Sau khi thu thập dữ liệu, chúng tôi tiến hành xử lý dữ liệu để chuẩn bị cho việc phân tích và trực quan hóa.
4. Phân tích dữ liệu: Sử dụng các hàm và gói phân tích dữ liệu trong R để thực hiện các phân tích cần thiết theo yêu cầu của đồ án.
5. Trực quan hóa dữ liệu: Sử dụng các gói trực quan hóa dữ liệu trong R như ggplot2, plotly để tạo các biểu đồ và đồ thị trực quan hóa kết quả phân tích.
6. Viết báo cáo: Cuối cùng, chúng tôi viết báo cáo về kết quả phân tích và trực quan hóa dữ liệu, mô tả quá trình thực hiện và kết luận của đồ án.

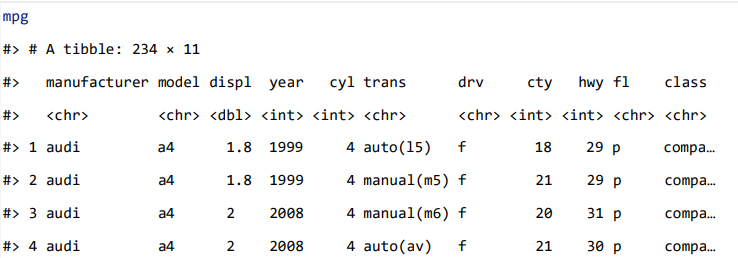
# **CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ THƯ VIỆN GGPLOT VÀ PHÂN TÍCH THỐNG KÊ**

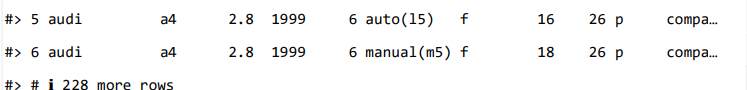
## **1.1. THĂM DÒ DỮ LIỆU BẰNG BIỂU ĐỒ GGPLOT2**

****

*Hình 1.1: Minh họa quy trình thêm các lớp hàm phân tích trong ggplot*

Để minh họa cho việc sử dụng ggplot chúng ta sẽ làm việc trên một dữ liệu tích hợp cùng ggplot2, đó là dữ liệu mpg chứa các quan sát được Cơ quan Bảo vệ Môi trường Hoa Kỳ thu thập trên 38 mẫu ô tô với 233 quan sát và 11 biến. (trong thư viện ggplot2 thuộc R gõ: ?mpg để biết chi tiết về nguồn gốc dữ liệu.

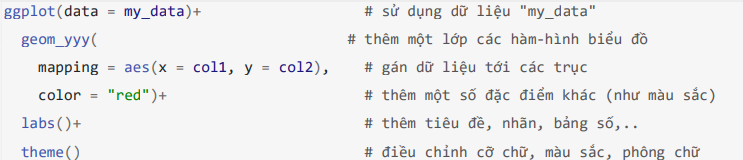
****

****

*Bảng 1.1: Dữ liệu quan sát về các mẫu ô tô (nguồn: ggplot2)*

*a) Cú pháp cơ bản*

Chúng ta có thể minh họa cú pháp cơ bản như sau:

****

*Hình 1.2: Cách gán biến số vào các trục*

*b) Gán các biến dữ liệu cho biểu đồ*

Hầu hết các hàm-hình geom phải được cho biết cái gì được sử dụng để vẽ biểu đồ, vì vậy chúng ta phải cung cấp cách map (gán) các biến số trong dữ liệu tới các thành phần của biểu đồ như là các trục, màu đối tượng, kích thước đối tượng, v.v. Đối với hầu hết các geoms, các thành phần thiết yếu phải được gán tới các cột trong dữ liệu là trục x, và (nếu cần) là trục y.

*c) Tính thẩm mỹ trong biểu đồ*

Tính thẩm mỹ trong biểu đồ có thể là màu sắc, kích thước, độ trong suốt, vị trí, v,v của dữ liệu liệu được vẽ. Không phải tất cả các geoms sẽ có các tùy chọn về tính thẩm mỹ, trang trí giống nhau, nhưng một số tùy chọn được áp dụng với phần lớn các geoms. Dưới đây là một số trang trí hay gặp:

* shape = Hiển thị một điểm với hàm geom\_point() dưới dạng dấu chấm, ngôi sao, hình tam giác hoặc hình vuông,...
* fill = Màu sắc bên trong (vd: của cột hoặc boxplot).
* color = Đường bên ngoài của cột, boxplot, v.v., hoặc màu của điểm nếu sử dụng hàm geom\_point().
* size = Kích thước (vd: độ dày của đường, kích thước của điểm).
* alpha = Độ trong suốt (1 = bình thường, 0 = vô hình).
* binwich = Độ rộng các bins trong biểu đồ histogram.
* width = Độ rộng của các cột trong “biểu đồ cột”.
* linetype = Kiểu của đường (vd: liền, nét đứt, chấm chấm).

Trang trí của đối tượng biểu đồ có thể được gán giá trị theo hai cách: Gán một giá trị tĩnh(vd: color = “blue”) để áp dụng cho tất cả các quan sát được vẽ biểu đồ hoặc gán cho từng biến của dữ liệu (vd: color = hospital) để hiển thị từng quan sát phụ thuộc vào giá trị của nó trong biến đó.

* Trang trí với một giá trị tĩnh

Nếu muốn yếu tố trang trí cho đối tượng biểu đồ tĩnh, nghĩa là - giống nhau đối với mọi quan sát trong dữ liệu, chúng ta gán nó bên trong geom nhưng ở bên ngoài đối với số mapping = aes(). Các phép gán này có thể ví dụ như: size = 1 hoặc color = “blue”.

* Trang trí theo giá trị của từng biến

Để thực hiện được điều này, chúng ta gán yếu tố trang trí của biển đồ với một biến ( không trong dấu ngoặc kép). Điều này phải được thực hiện bên trong một hàm mapping = aes().

* Trang trí theo nhóm đối tượng

Lưu ý rằng tùy thuộc vào loại geom sử dụng, chúng ta sẽ cần sử dụng các đối số khác nhau để trang trí cho nhóm đối tượng. Đối với geom\_point(), ta thường sử dụng các tham số như color , shape hoặc size. Trong khi đó đối với geom\_bar(), ta thường sử dụng tham số fill. Điều này chỉ phụ thuộc vào loại geom và yếu tố trang trí nào mà chúng ta muốn thể hiện sự phân nhóm.

*d) Gán nhãn cho biểu đồ*

Việc đặt tên cho tiêu đề biểu đồ, tên các biến trên trục, các chú thích là công việc không thể thiếu khi vẽ biểu đồ, và việc này được thực hiện với hàm labs() bằng cách thêm dấu + như cách chúng ta thêm các geoms.

Bên trong hàm labs(), cung cấp các chuỗi ký tự cho các đối số sau:

▪ x = và y = Tiêu đề trục x và trục y (nhãn).

▪ title = Tiêu đề chính của biểu đồ.

▪ subtitle = Tiêu đề phụ của biểu đồ, nhỏ hơn và đặt bên dưới tiêu đề chính.

▪ caption = Chú thích của biểu đồ, mặc định ở góc phải dưới.

*e) Căn chỉnh trong biểu đồ*

Việc căn chỉnh màu nền của biểu đồ, sự xuất hiện/biến mất của đường lưới, cũng như phông chữ/cỡ chữ/màu sắc/căn lề của văn bản (tiêu đề chính, tiêu đề phụ, Chú thích, chữ trên các trục…). được thực hiện theo hai cách: Căn chỉnh theo mặc định sẵn có và căn chỉnh cá nhân đơn lẻ

* Căn chỉnh theo mặc định

Căn chỉnh theo mặc định, tức là chúng ta sử dụng căn chỉnh theo một chủ đề hoàn chỉnh bằng hàm theme\_() để điều chỉnh toàn bộ các thành phần biểu đồ. Cách căn chỉnh này khá đơn giản, chúng ta có thể sử dụng một số hàm chủ đề hoàn chỉnh bên dưới đây.

o theme\_gray(): Chủ đề ggplot2 đặc trưng với nền màu xám và đường lưới màu trắng, được thiết kế để đưa dữ liệu về phía trước nhưng vẫn giúp việc so sánh trở nên dễ dàng.

o theme\_bw(): Chủ đề ggplot2 tối trên ánh sáng cổ điển. Có thể hoạt động tốt hơn cho bài thuyết trình trình chiếu bằng máy chiếu.

o theme\_linedraw(): Một chủ đề chỉ có các đường màu đen có chiều rộng khác nhau trên nền trắng, gợi nhớ đến một bản vẽ đường. Phục vụ mục đích tương tự như theme\_bw(). Lưu ý rằng chủ đề này có một số dòng rất mỏng (<< 1 pt) khi in ấn rất dễ mất hình ảnh.

o theme\_light(): Một chủ đề tương tự như theme\_linedraw() nhưng có các đường và trục màu xám nhạt, để hướng sự chú ý nhiều hơn tới dữ liệu.

o theme\_dark(): Tương tự màu tối của theme\_light(), với kích thước dòng tương tự nhưng nền tối, hữu ích để làm nổi bật những đường màu mảnh.

o theme\_minimal(): Một chủ đề tối giản không có chú thích nền.

o theme\_classic(): Một chủ đề có giao diện cổ điển với các đường trục x và y và không có đường lưới.

o theme\_void(): Một chủ đề hoàn toàn trống rỗng.

o theme\_test(): Một chủ đề cho bài kiểm tra đơn vị trực quan. Lý tưởng nhất là nó không bao giờ thay đổi ngoại trừ cho các tính năng mới.

* Căn chỉnh cá nhân đơn lẻ

Hàm theme() có thể nhận một số lượng lớn các đối số, mỗi đối số sẽ chỉnh sửa một khía cạnh rất cụ thể của biểu đồ. Chúng ta sẽ không trình bày tất cả các đối số, nhưng sẽ tập trung mô tả công thức chung cho chúng và chỉ cách tìm tên đối số khi cần. Cú pháp cơ bản là:

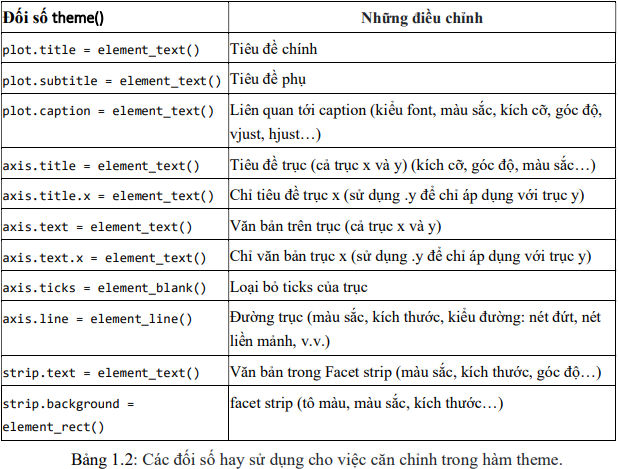
o Bên trong hàm theme(), hãy viết tên đối số cho phần tử biểu đồ mà ta muốn chỉnh sửa, chẳng hạn như plot.title =

o Cung cấp một hàm element\_() tới đối số.

o Thường sử dụng nhất là element\_text(), một số khác bao gồm element\_rect() chọn màu nền cho canvas, hoặc element\_blank() để xóa các phần tử biểu đồ.

o Bên trong hàm element\_(), xác định giá trị đối số cần gán để điều chỉnh theo ý bạn mong muốn.

Sau đây là một số đối số phổ biến của hàm theme().



*Bảng 1.2: Các đối số hay sử dụng cho việc căn chỉnh trong hàm theme.*

Có một số đối số khác ít phổ biến hơn, nhưng nếu cần chúng ta có thể liệt kê ra chúng bằng cách: Chạy lệnh theme\_get() từ ggplot2 để in tất cả hơn 90 đối số của hàm theme() ra console. Hoặc nếu chúng ta muốn xóa một phần tử của biểu đồ, bạn cũng có thể làm điều đó bằng hàm theme(). Chỉ cần đặt element\_blank() tới đối số để nó biến mất hoàn toàn. Đối với chú thích, thiết lập legend.position = "none".

*f) Phối màu sắc, tô màu, thang đo*

* Phối màu

Để phối màu sắc của các đối tượng biểu đồ (geoms/shapes) ví dụ như điểm, cột, đường, ô, v.v. chúng ta sẽ điều chỉnh color = (màu bên ngoài) hoặc fill = (màu bên trong), riêng đối với geom\_point(), ta chỉ có thể điều khiển color =, để xác định màu của điểm. Khi thiết lập màu hoặc tô màu, chúng ta có thể sử dụng tên màu được R nhận dạng như "red" (xem danh sách các màu đầy đủ gõ ?colors trong cửa sổ soạn thảo hoặc ấn F1).

* Thang đo cho yếu tố trang trí (thẩm mỹ)

Khi gán một biến với một yếu tố thẩm mỹ của biểu đồ (vd: x =, y =, fill =, color =…), biểu đồ sẽ hiển thị một thang đo/chú giải, trên đó có thể là các giá trị liên tục, rời rạc, ngày tháng, v.v. tùy thuộc vào kiểu dữ liệu của biến được chỉ định. Nếu ta có nhiều yếu tố thẩm mỹ được gán tới biến, biểu đồ sẽ có nhiều thang đo.

Chúng ta có thể kiểm soát các thang đo bằng hàm scales\_() thích hợp. Các hàm scale của ggplot() có 3 phần được viết như sau: scale\_aesthetic\_method().

o Phần đầu tiên, scale\_(), là cố định.

o Phần thứ hai, aesthetic, là tên yếu tố thẩm mỹ bạn muốn điều chỉnh thang đo (\_fill\_, \_shape\_, \_color\_, \_size\_, \_alpha\_…). Các tùy chọn ở đây cũng bao gồm \_x\_ và \_y\_.

o Phần thứ ba, method, sẽ là một trong số các tùy chọn sau \_discrete(), continuous(), \_date(), \_gradient(), hoặc \_manual(), tùy thuộc vào kiểu dữ liệu của biến và cách chúng ta muốn kiểm soát nó. Có những tùy chọn khác, tuy nhiên những lựa chọn trên thường được sử dụng nhất.

* Các đối số của hàm Scale

Mỗi loại thang đo có những đối số riêng của chúng, mặc dùcũng có những sự trùng nhau (Truy vấn hàm chẳng hạn như ?scale\_color\_discrete trong cửa sổ R console để xem tài liệu về các đối số của hàm).

Với thang đo liên tục, sử dụng breaks = để cung cấp một chuỗi giá trị tới seq() (đặt to =, from =, và by = ). Thiết lập expand = c(0,0) để loại bỏ không gian đệm xung quanh các trục (điều này có thể được sử dụng trên bất kỳ thang đo của trục \_x\_ hoặc \_y\_ ).

Với thang đo rời rạc, ta có thể điều chỉnh thứ tự của các giá trị với breaks =, và cách các giá trị hiển thị với đối số labels =, cung cấp một vectơ ký tự cho mỗi cái đó. Chúng ta cũng có thể loại bỏ NA dễ dàng bằng cách đặt na.translate = FALSE.

* Điều chỉnh thủ công

Chúng ta có thể sử dụng các hàm scaling “một cách thủ công” để gán màu sắc như mong muốn.

o Gán màu cho các giá trị dữ liệu với đối số values =.

o Cụ thể màu sắc cho giá trị NA với na.value =.

o Thay đổi cách các giá trị được viết trong chú giải với đối số labels =.

o Thay đổi tiêu đề chú giải bằng name =.

* Thang đo trên các trục

Khi dữ liệu được ánh xạ tới các trục của biểu đồ, chúng cũng có thể được điều chỉnh bằng các lệnh scales. Phổ biến là điều chỉnh hiển thị của một trục (ví dụ: trục y) được ánh xạ tới một biến có dữ liệu liên tục.

Chúng ta có thể điều chỉnh độ chia hoặc hiển thị của giá trị trong ggplot bằng cách sử dụng scale\_y\_continuous(). Như đã lưu ý ở trên, sử dụng đối số breaks = để cung cấp một chuỗi các giá trị sẽ đóng vai trò là “ngắt các khoảng giá trị” dọc theo thang đo. Đây là những giá trị mà các số sẽ hiển thị. Đối với đối số này, ta có thể cung cấp một vectơ c() chứa các giá trị để chia thang đo theo mong muốn hoặc bạn có thể cung cấp một chuỗi số thông thường bằng cách sử dụng hàm seq() từ base R. Hàm seq() này chấp nhận to =, from =, và by =.

* Hiển thị phần trăm trên trục

Nếu giá trị dữ liệu ban đầu là tỷ lệ, chúng ta có thể hiển thị chúng dưới dạng phần trăm với “%” bằng cách cung cấp labels = scales::percent trong lệnh scales command. Ngoài ra, có một giải pháp thay thế là chuyển đổi các giá trị thành ký tự và thêm ký tự “%” vào cuối, cách tiếp cận này sẽ gây ra phức tạp vì dữ liệu sẽ không còn là các giá trị số liên tục.

* Thang đo log

Một số dữ liệu khi hiển thị trên biểu đồ có khoảng cách (metric) khá lớn, dẫn tới khó quan sát hoặc dữ liệu biểu diễn vượt ra ngoài khung hình của biểu đồ. Khi đó việc biến đổi một trục liên tục sang thang đo log sẽ khắc phục được những hạn chế này. Cách chuyển rất đơn giản bằng cách thêm trans = "log2" vào lệnh scale.

* Thang đo Gradient

Tô màu theo thang đo gradient liên quan đến bản nhiệt. Các giá trị mặc định thường khá dễ chịu, nhưng ta có thể muốn điều chỉnh các giá trị, điểm cắt, v.v.

*g) Lưu trữ, chỉnh sửa và xuất bản biểu đồ*

* Lưu biểu đồ

Mặc định khi chạy lệnh ggplot(), biểu đồ sẽ được in ở cửa số Plots của RStudio. Tuy nhiên, bạn cũng có thể lưu biểu đồ dưới dạng một đối tượng bằng cách sử dụng toán tử gán <- và đặt tên cho nó. Biểu đồ sẽ không được in ra trừ khi ta gọi tên của đối tượng. Ta cũng có thể in nó bằng cách đưa tên biểu đồ vào hàm print(), nhưng điều này chỉ cần thiết trong một số trường hợp nhất định chẳng hạn như khi biểu đồ được tạo bên trong một vòng lặp for để in nhiều biểu đồ cùng một lúc.

* Chỉnh sửa biểu đồ đã lưu

Một điểm hay của ggplot2 là ta có thể gán tên cho một biểu đồ (như bên trên), và sau đó thêm các lớp mới hoặc chỉnh sửa bắt đầu bằng tên của nó. Chúng ta không cần phải lặp lại tất cả các lệnh đã tạo ra biểu đồ ban đầu.

* Xuất bản biểu đồ

Việc xuất bản biểu đồ được thực hiện dễ dàng với hàm ggsave() của package ggplot2 hoặc chức năng Export trong Rstudio.

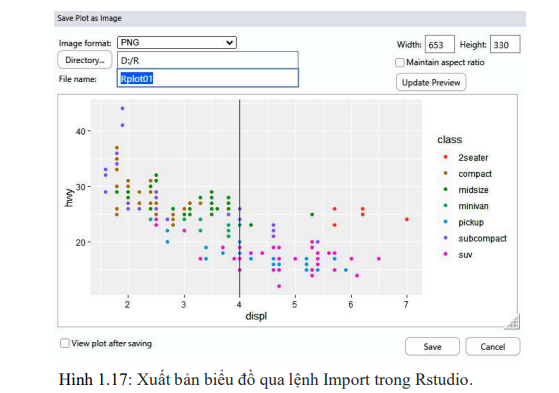
o Với hàm ggsave(), có thể được tiến hành theo hai cách:

▪ Chỉ định tên của đối tượng biểu đồ, sau đó là đường dẫn tệp và tên có phần mở rộng. Ví dө: ggsave(Bieu\_do1, here("plots", " Bieu\_do1.png")).

▪ Chạy lệnh chỉ với một đường dẫn tệp, để lưu biểu đồ gần nhất được in ra.Ví dụ: ggsave(here("plots", " Bieu\_do1.png")).

Chúng ta có thể xuất dưới dạng tệp png, pdf, jpeg, tiff, bmp, svg, hoặc một số định dạng khác, bằng cách chỉ định phần mở rộng tệp trong đường dẫn tệp. Hơn nữa, ta cũng có thể chỉ định các đối số width =, height =, và units = (“in”, “cm”, hoһc “mm”), và chỉ định dpi = để điều chỉnh độ phân giải của biểu đồ (vd: dpi = 300). Xem hướng dẫn chi tiết về hàm bằng cách gõ ?ggsave trong Rstudio.

o Với Export trong Rstudio, chúng ta có thể lựa chọn save image; pdf hoặc copy to clipboard,… Khi chọn save image chúng ta sẽ có 1 bảng thông số như hình dưới đây:



*Hình 1.3:Xuất bản biểu đồ qua lệnh Import trong Rstudio*

Chúng ta điền các đối số width =, height =,…. phù hợp với mục đích sử dụng.

## **1.2. TÓM TẮT KẾT QUẢ THEO SUY DIỄN THỐNG KÊ**

Tóm tắt các kết quả theo suy diễn thống kê như các tính toán về đặc trưng của dữ liệu mẫu, các bài toán ước lượng, bài toán kiểm định giả thuyết, bài toán phân tích hồi quy tạo thành các mô hình phân tích thống kê. Những module này kết hợp với phân tích dữ liệu qua biểu đồ sẽ cho kết quả trực quan dữ liệu chính xác hơn. Ngoài ra, trong phần này các ví dụ minh họa sẽ sử dụng file dữ liệu Diem\_TN, xem [1].

**1.2.1. Thống kê mô tả**

Cho một biến số  chúng ta có thể tính toán một số chỉ số thống kê mô tả như sau:

| Lý thuyết | Hàm R |
| --- | --- |
| Số trung bình: | mean (x) |
| Phương sai: | var (x) |
| Độ lệch chuẩn: | sd (x) |
| Trị số thấp nhất | min (x) |
| Trị số cao nhất | max (x) |
| Toản cự (range) | ranges (x) |

*Bảng 1.3: Các hàm tính thống kê mô tả cơ bản trong R*

### **1.2.2. Thống kê suy diễn trong các bài toán kiểm định**

*a) Trị số P-value:*

Trong nghiên cứu khoa học, ngoài những dữ kiện bằng số liệu, biểu đồ và hình ảnh, con số mà chúng ta thường gặp nhất là trị số P (P-value). Do đó, trước khi nói đến các phương pháp phân tích thống kê bằng R, chúng ta cùng tìm hiểu về ý nghĩa của trị số này.

Một giả thiết được xem là mang tính "khoa học" nếu giả thiết đó có khả năng "phản nghiệm". Theo Karl Popper, nhà triết học khoa học, đặc điểm duy nhất để phân biệt giữa một lý thuyết khoa học thực sự với một lý thuyết giả khoa học là lý thuyết khoa học luôn có tính chất có thể "bị bác bỏ" (hay bị phản bác – falsified) bằng những thử nghiệm đơn giản. Ông gọi đó là "khả năng phản nghiệm". Phép phản nghiệm là phương cách tiến hành những thử nghiệm không phải để xác minh mà để phê phán các lý thuyết khoa học và có thể coi đây như là một nền tảng cho khoa học thực sự. Có thể xem quy trình phản nghiệm là một cách học hỏi từ sai lầm. Khoa học phát triển cũng một phần là do học hỏi từ sai lầm.

Chúng ta có thể tóm tắt tiến trình của một nghiên cứu (dựa vào trị số P) như sau:

- Đưa ra một giả thiết chính (H1).

- Từ giả thiết chính, đưa ra một giả thiết đối (H0).

- Tiến hành thu thập dữ kiện (D).

- Phân tích dữ kiện: tính toán xác suất D xảy ra nếu H0 là sự thật. Nói theo ngôn ngữ toán học, bước này xác định P(D|H0).

Vì thế, giá trị P có nghĩa là xác suất của dữ kiện D xảy ra nếu giả thiết đối H0 là sự thật. Như vậy, giá trị P không trực tiếp cho chúng ta một ý niệm gì về sự thật của giả thiết chính H1; nó chỉ gián tiếp cung cấp bằng chứng để chúng ta chấp nhận giả thiết chính và loại bỏ giả thiết đảo.

*b) Các loại sai lầm trong kiểm định giả thiết:*

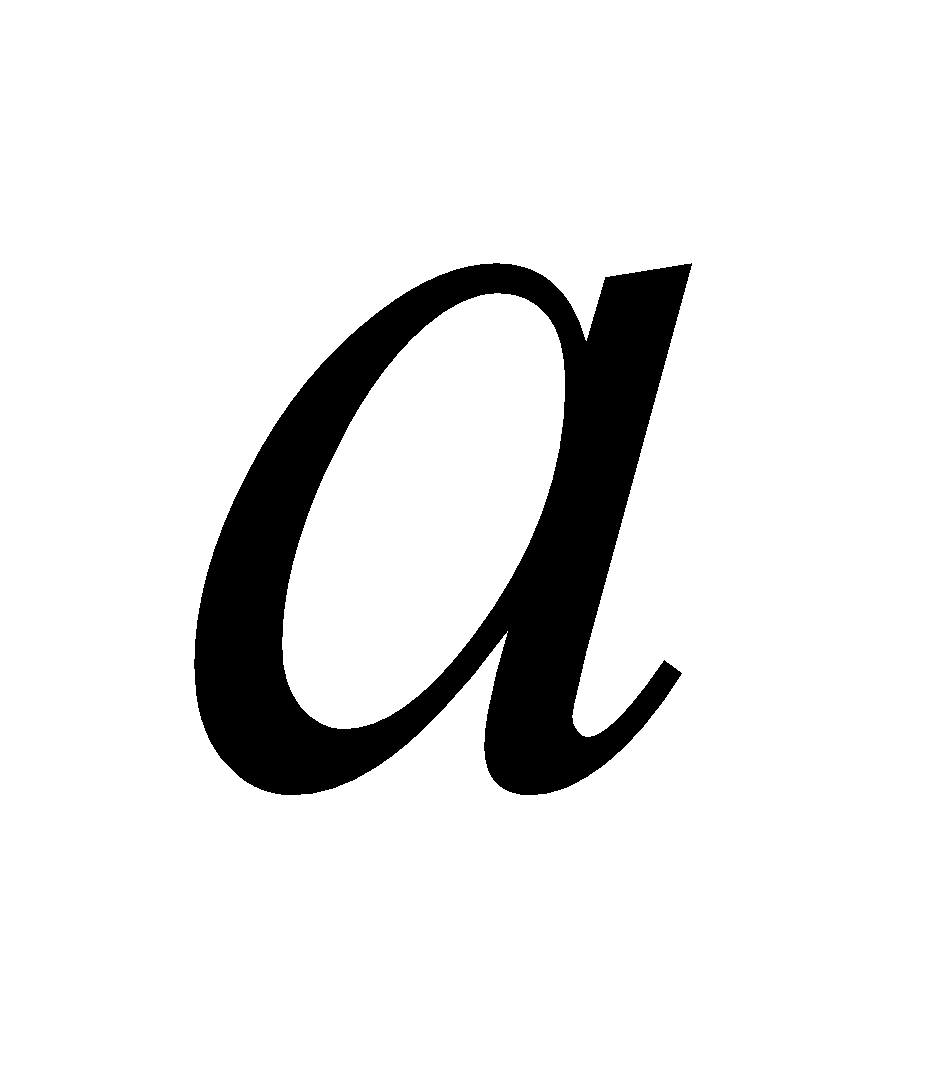
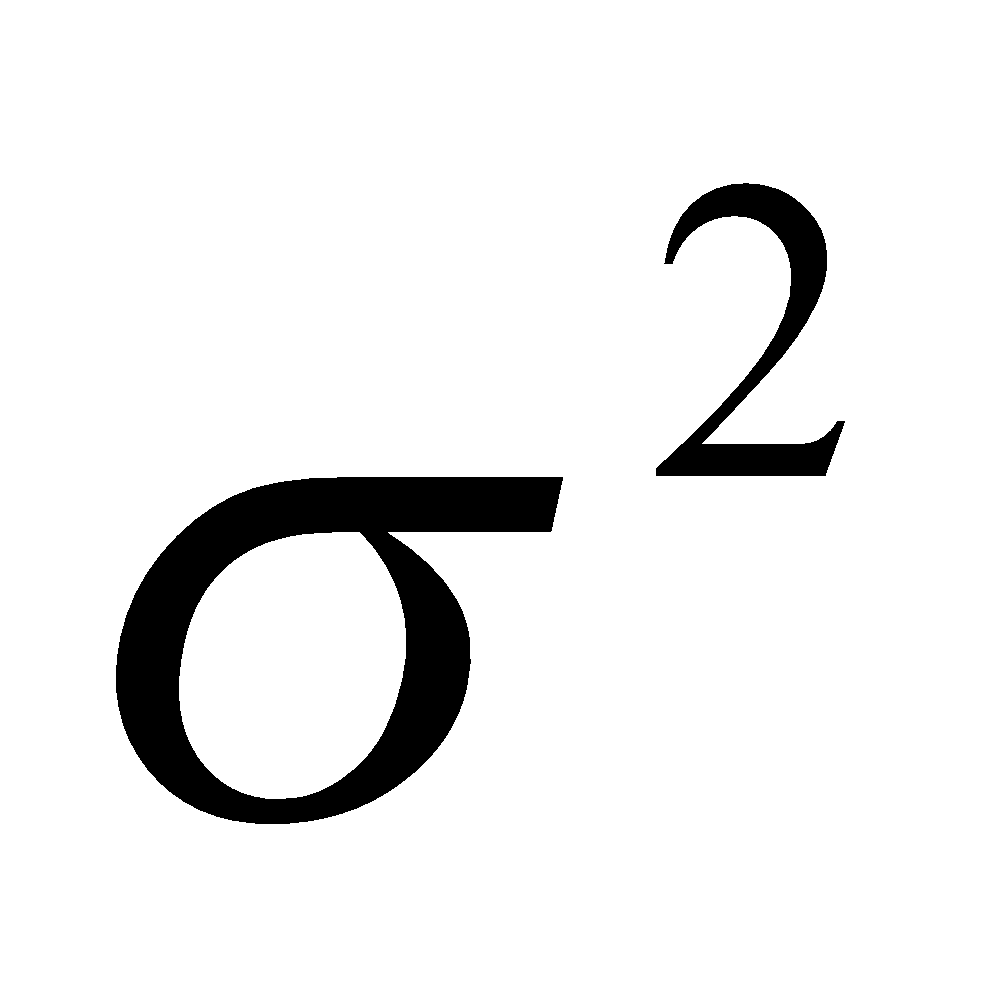
- Sai lầm loại I: Nếu ta loại bỏ H0 khi H0 đúng thì sai lầm đó được gọi là sai lầm loại I.

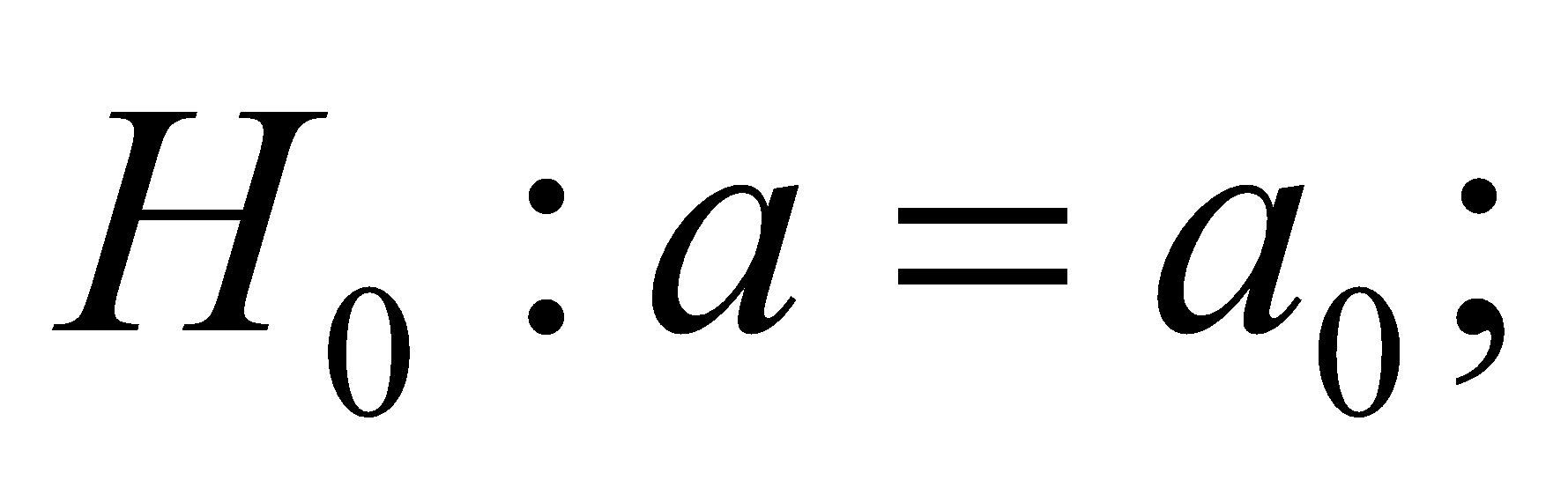
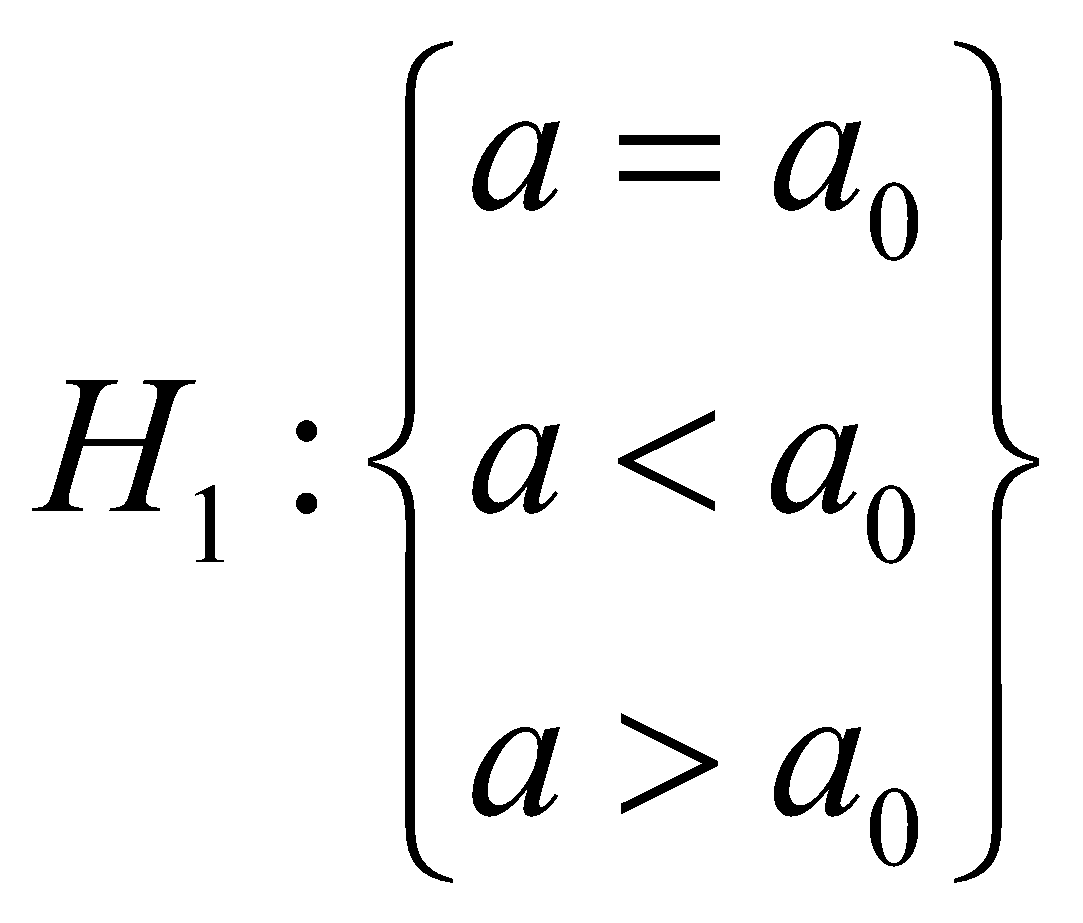
- Sai lầm loại II: Nếu H0 sai mà ta không loại bỏ H0 thì sai lầm đó được gọi là sai lầm loại II.

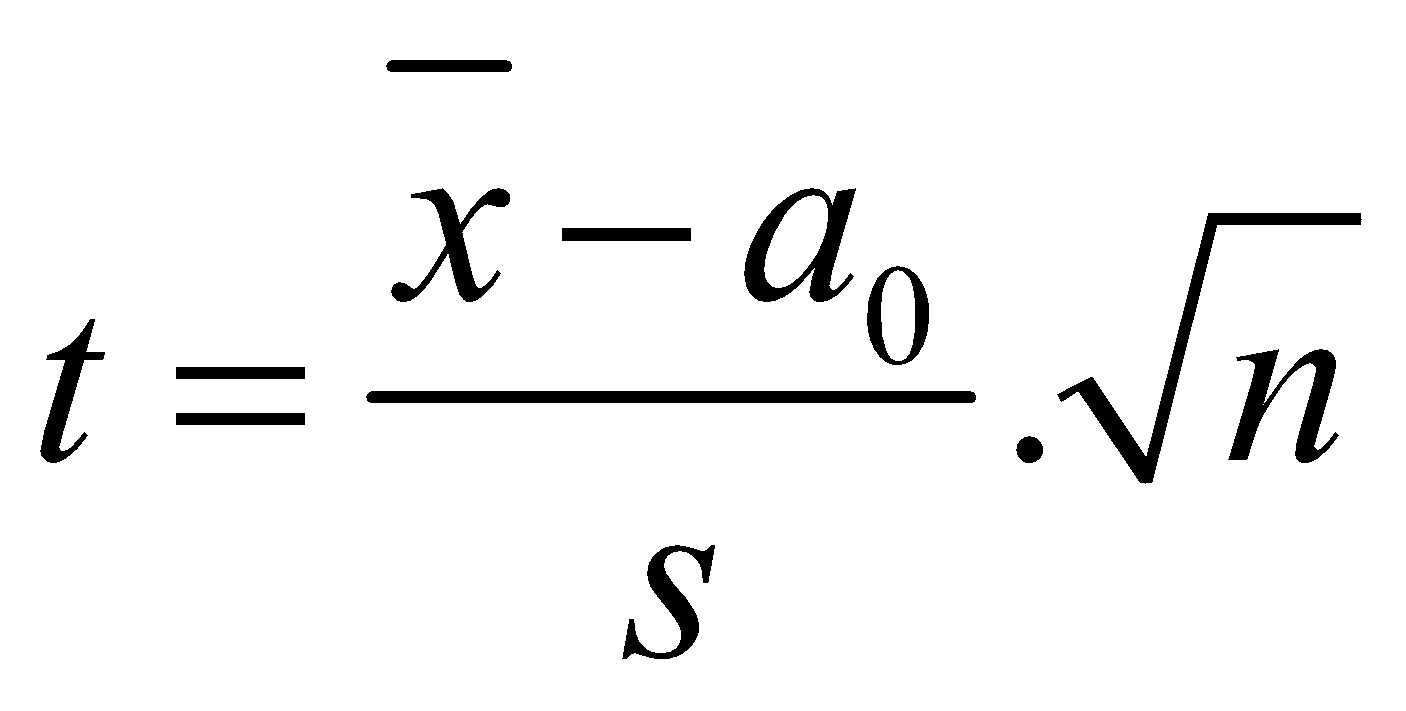
*c) Kiểm định t (t.test):*

Kiểm định t dựa vào giả thiết phân phối chuẩn. Có hai loại kiểm định t: kiểm định t cho một mẫu (one-sample t-test), và kiểm định t cho hai mẫu (two-sample t-test). Chúng ta sẽ minh họa hai kiểm định này thông qua dữ liệu của file Diem\_TN.

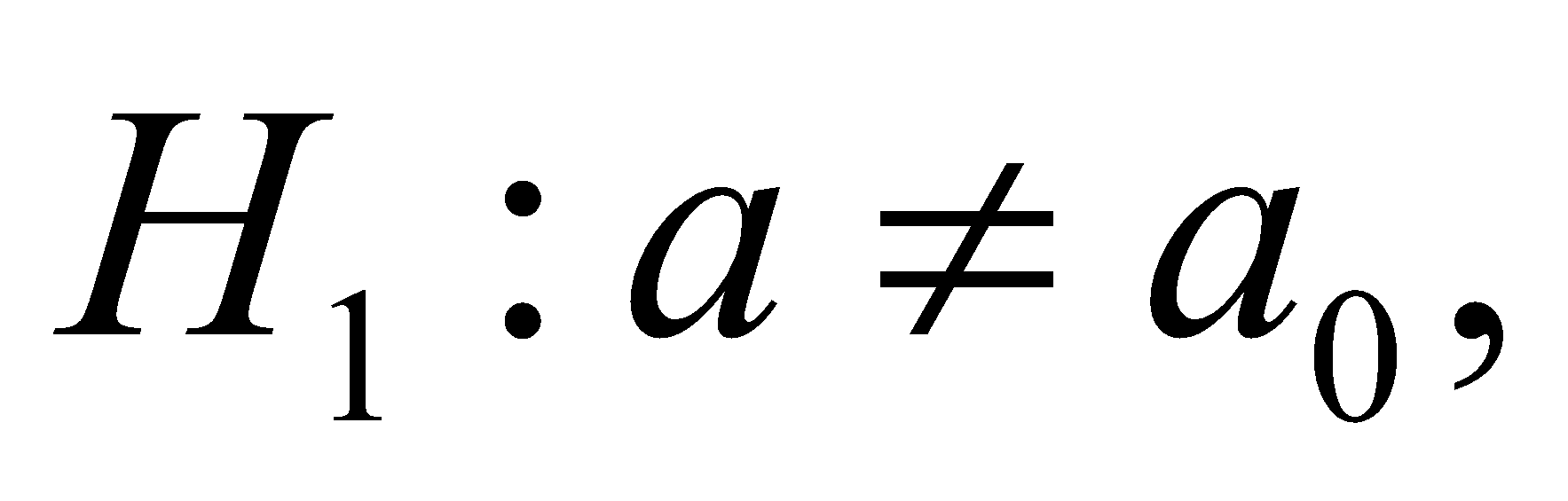
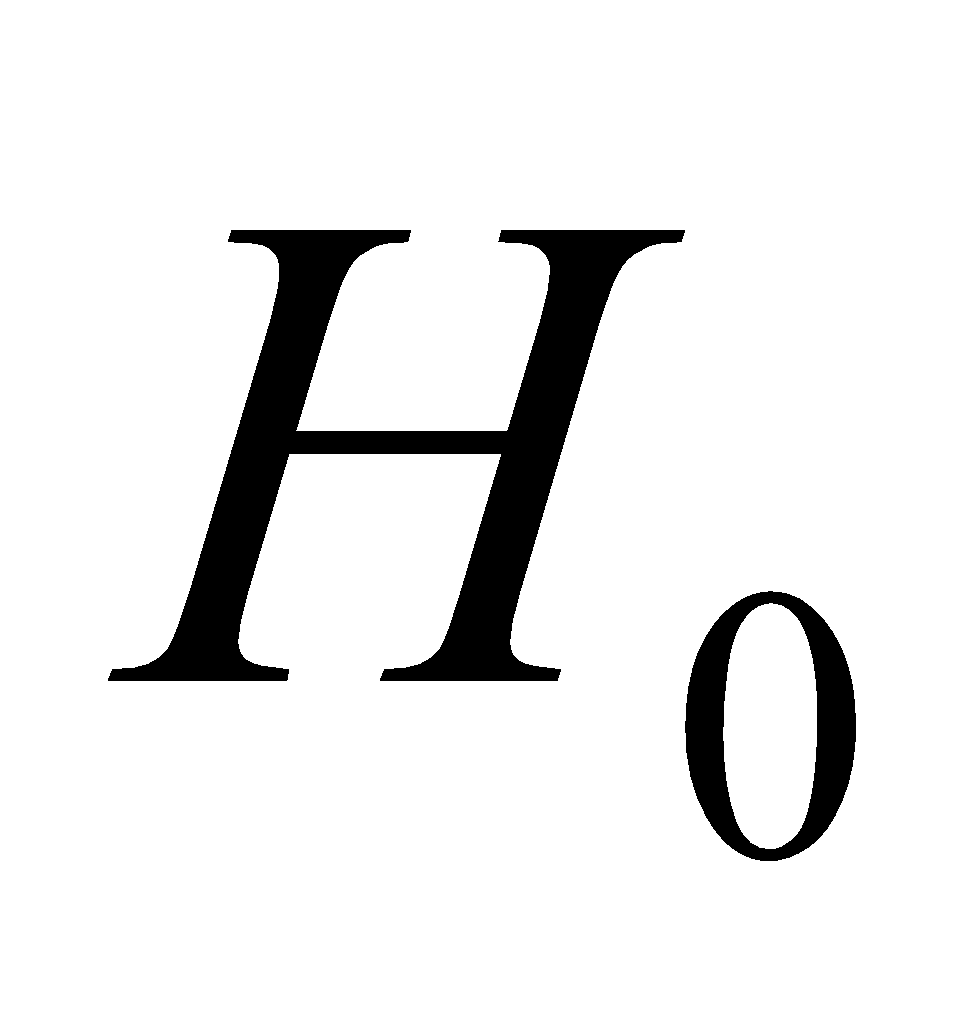
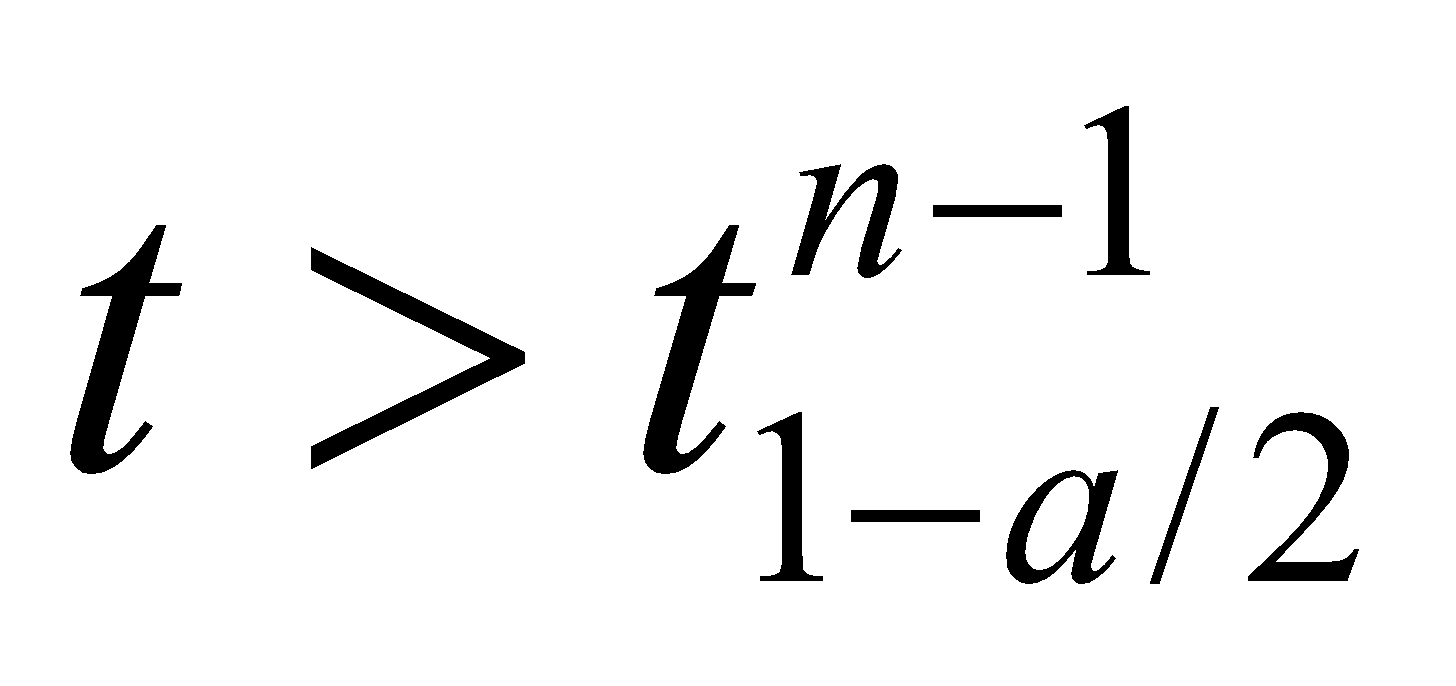
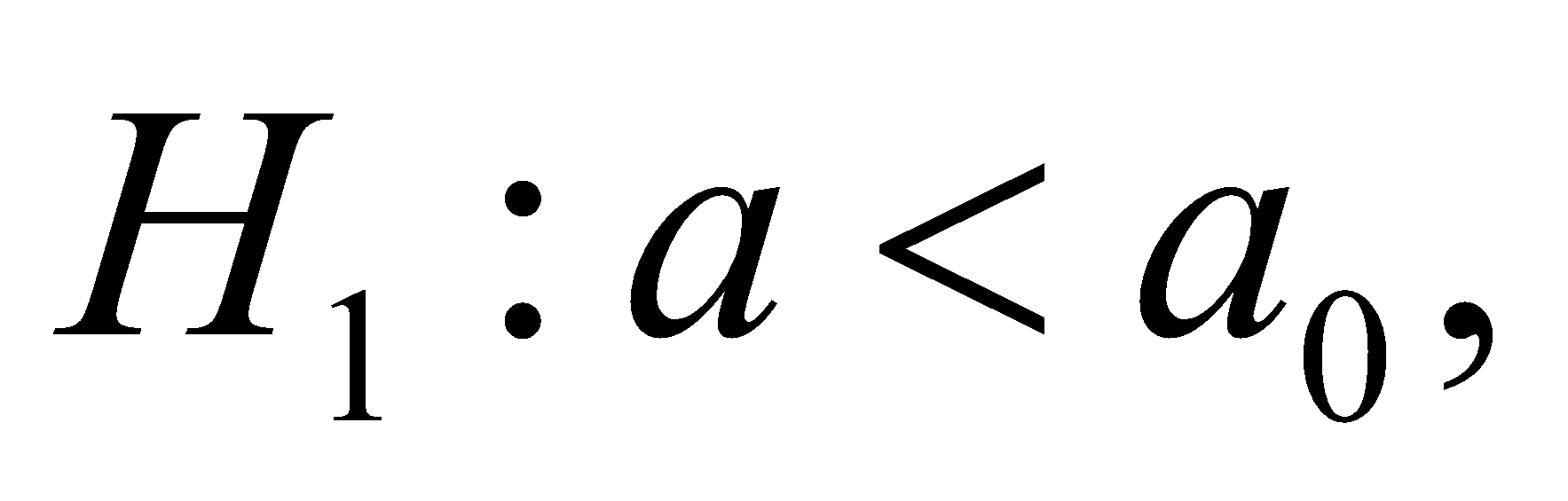
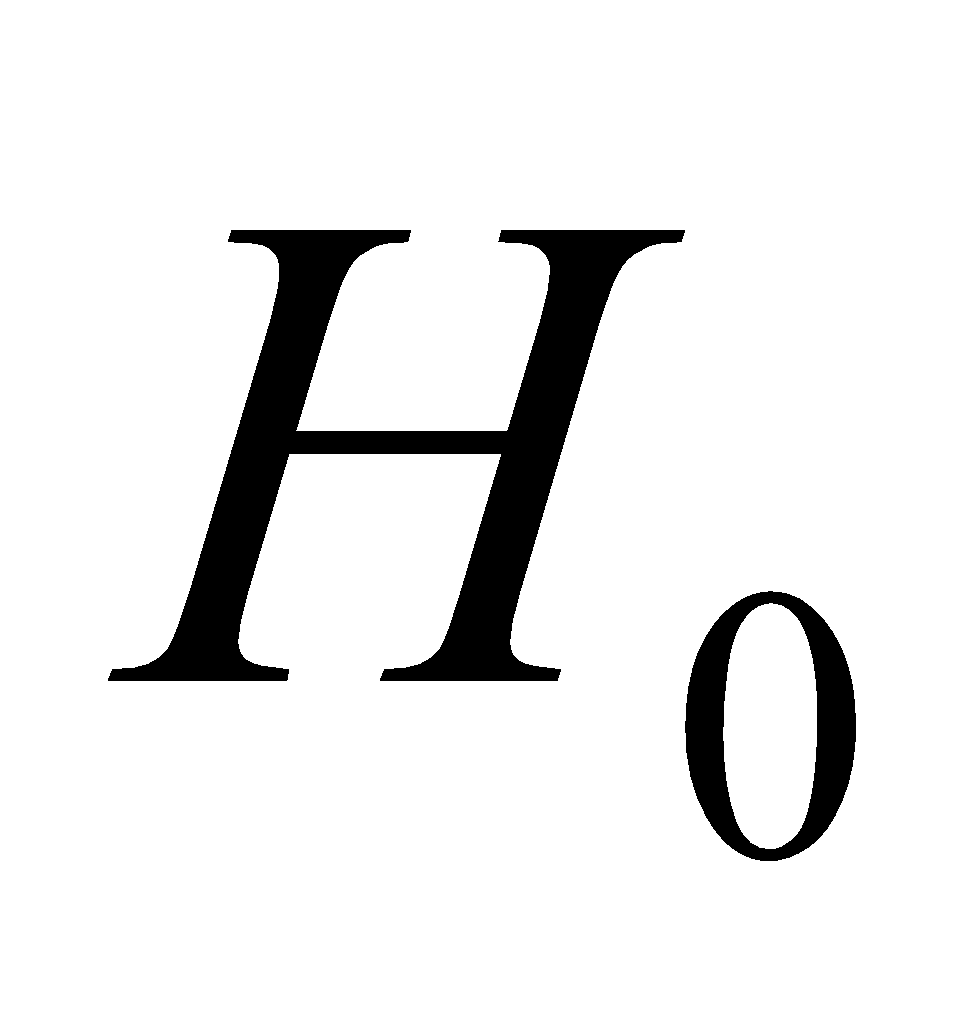
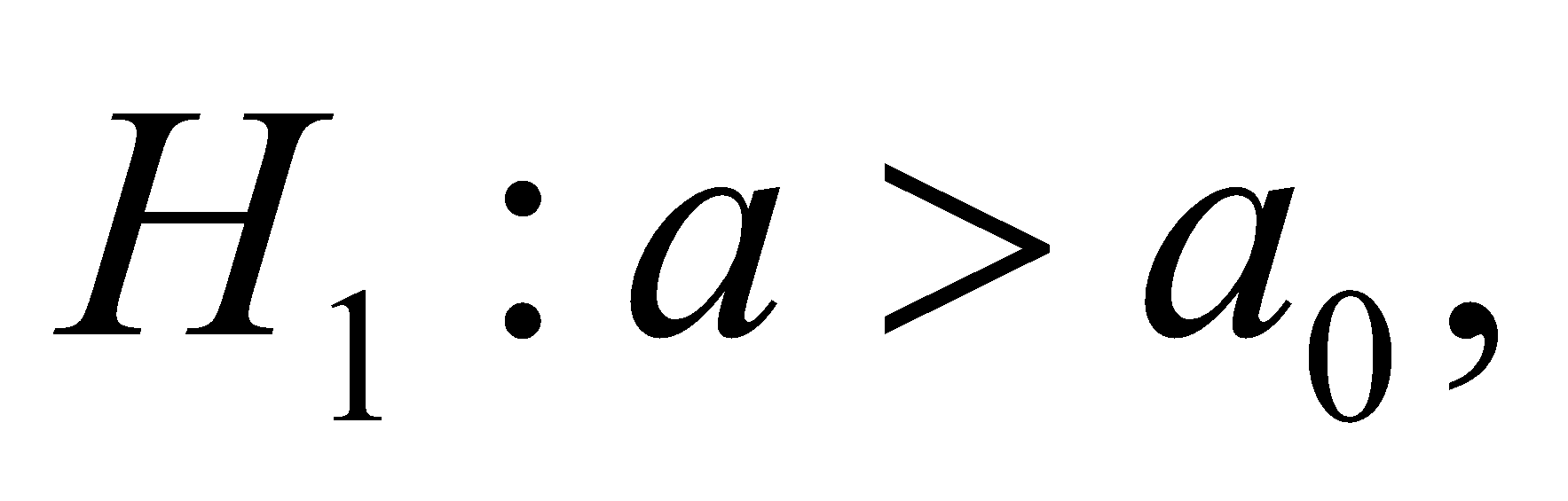
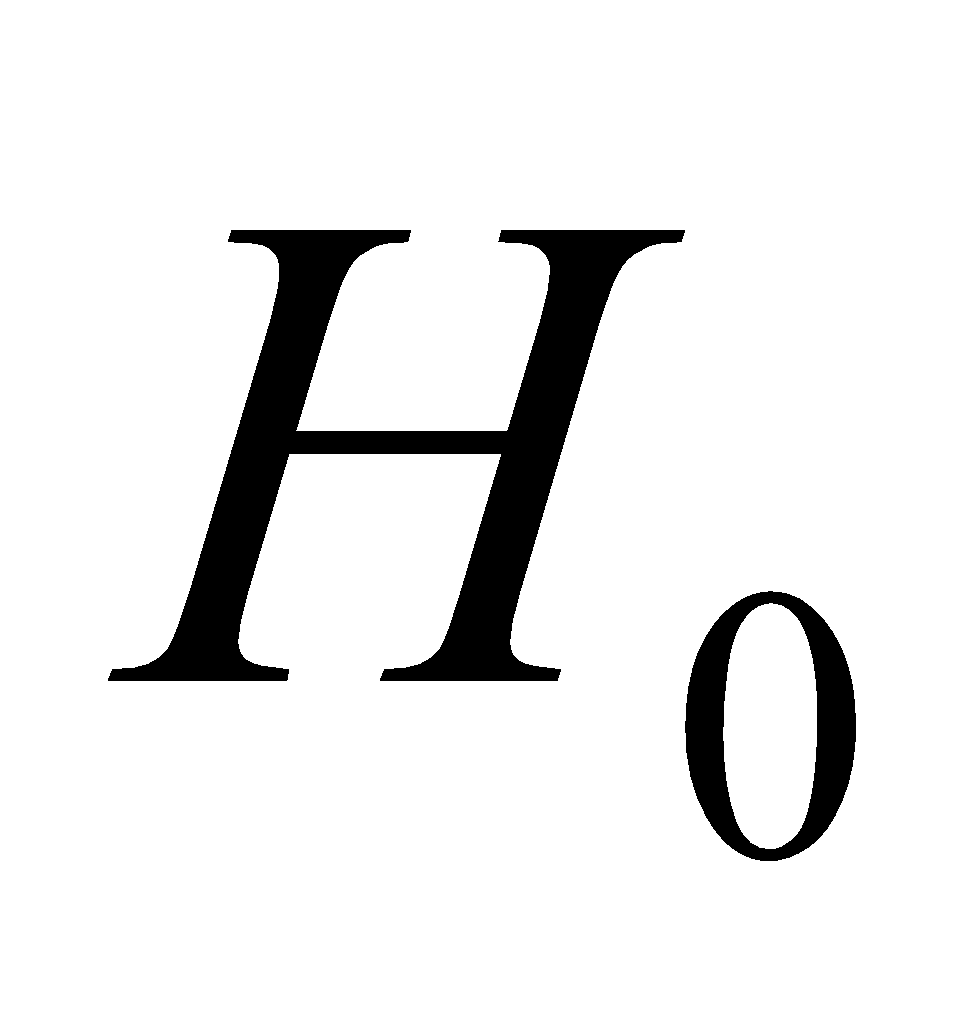
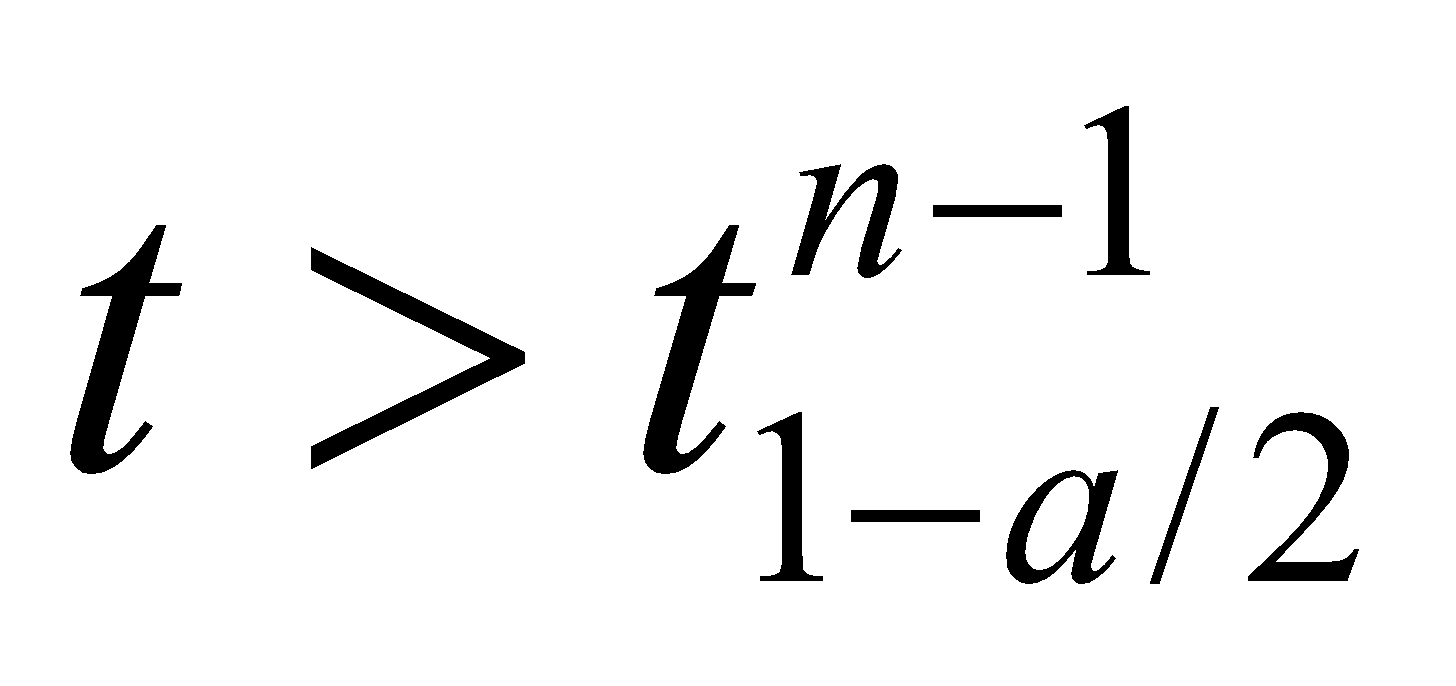
* **Kiểm định giả thuyết cho kỳ vọng một mẫu**

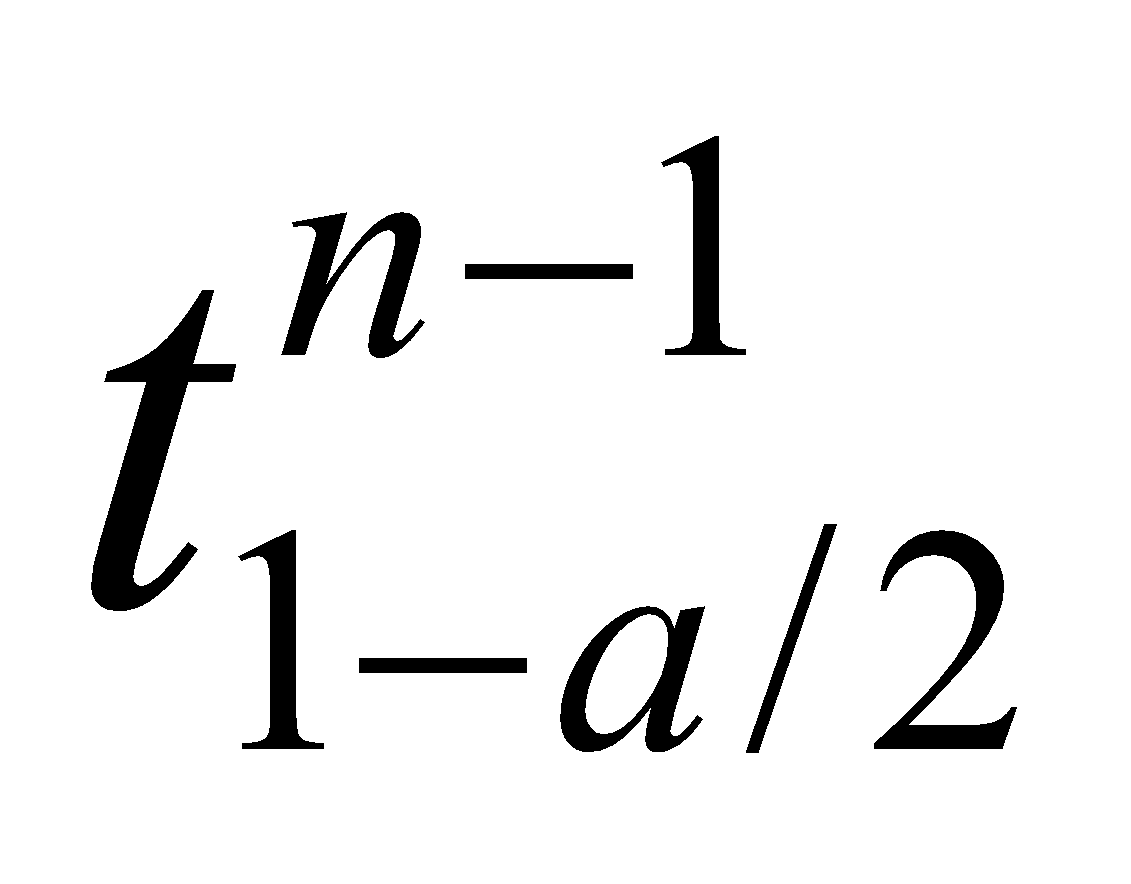
Xét mẫu ngẫu nhiên  được chọn từ tổng thể có phân phối chuẩn (hoặc xấp xỉ chuẩn tức phân phối có dạng đối xứng) với kỳ vọng và phương sai .

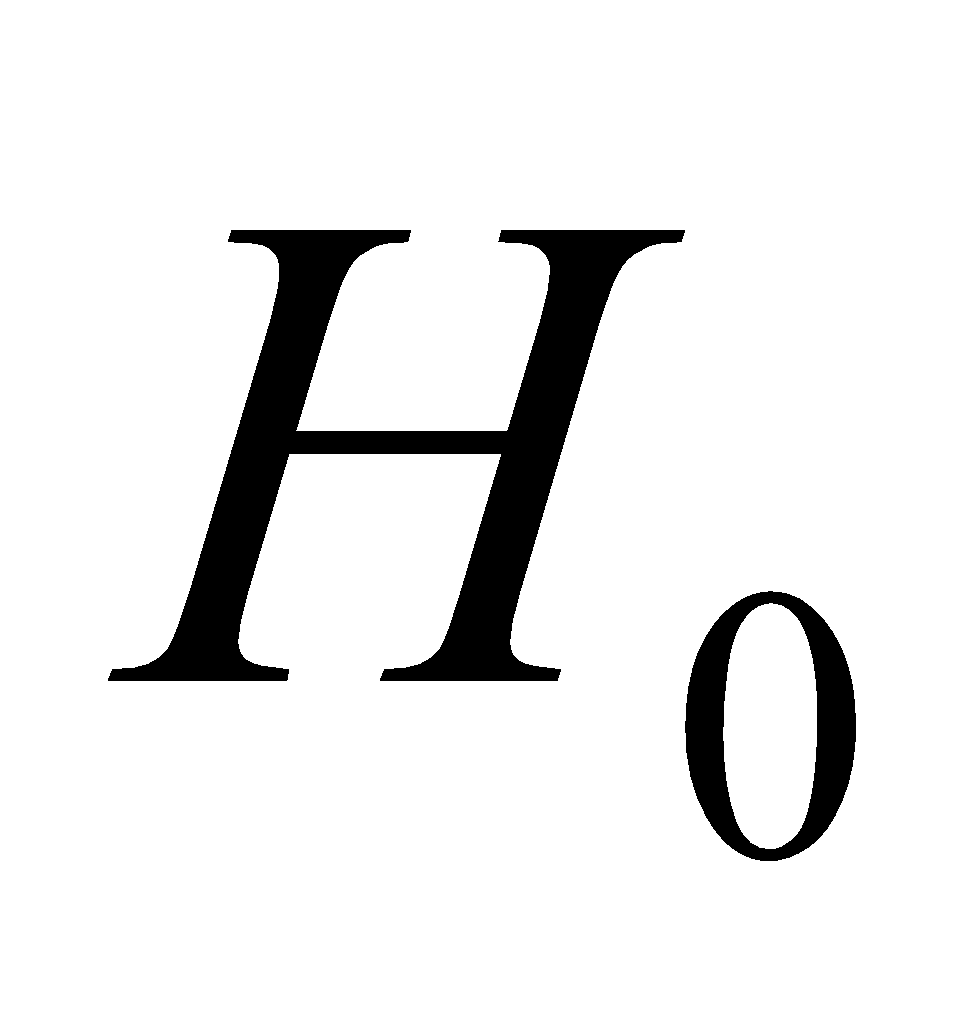
Giả thuyết  Đối thuyết (Một trong 3 trường hợp)

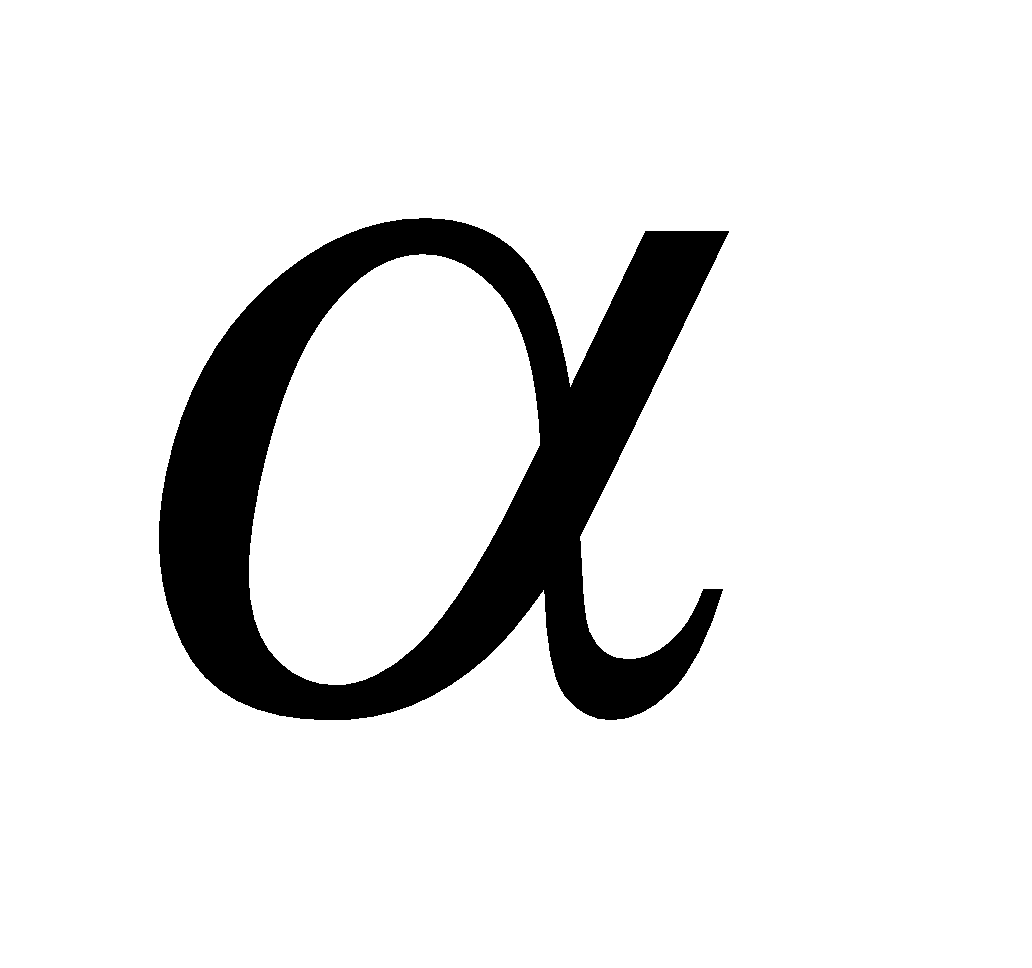
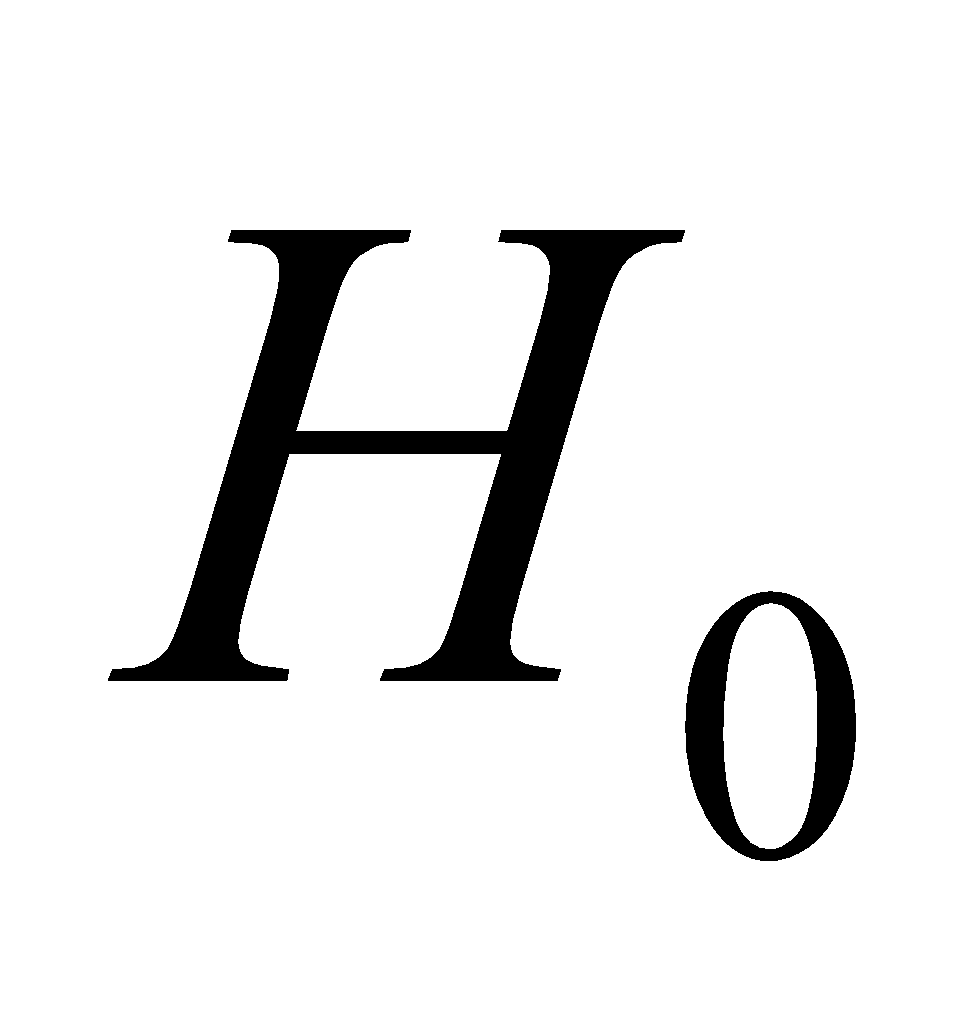
Tính thống kê kiểm định: .

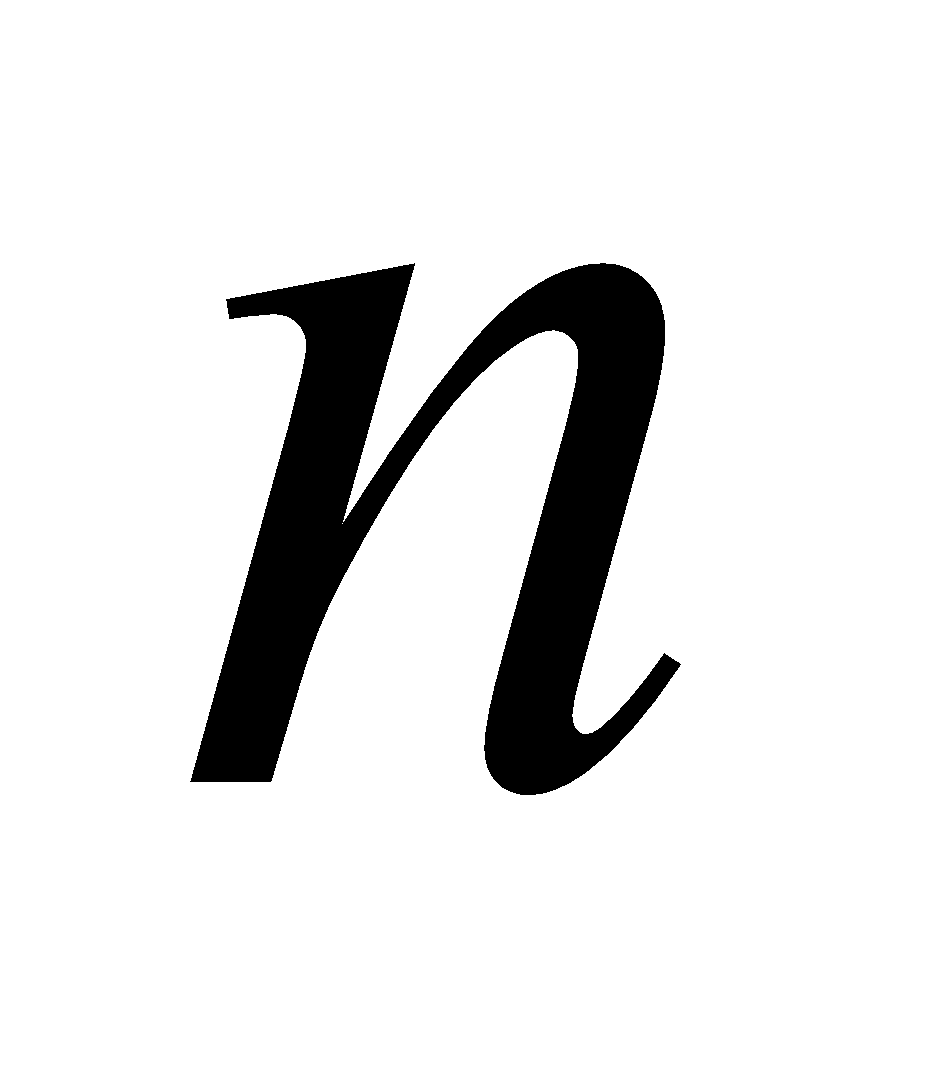
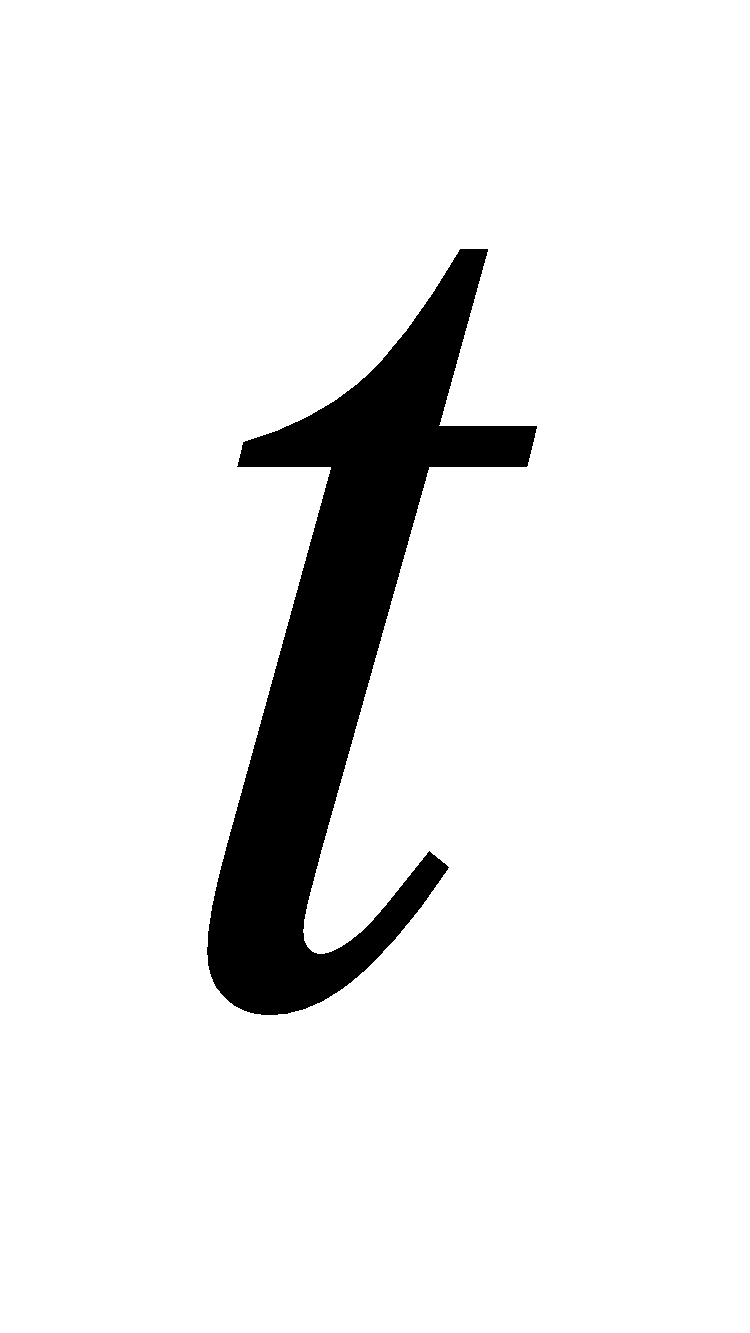
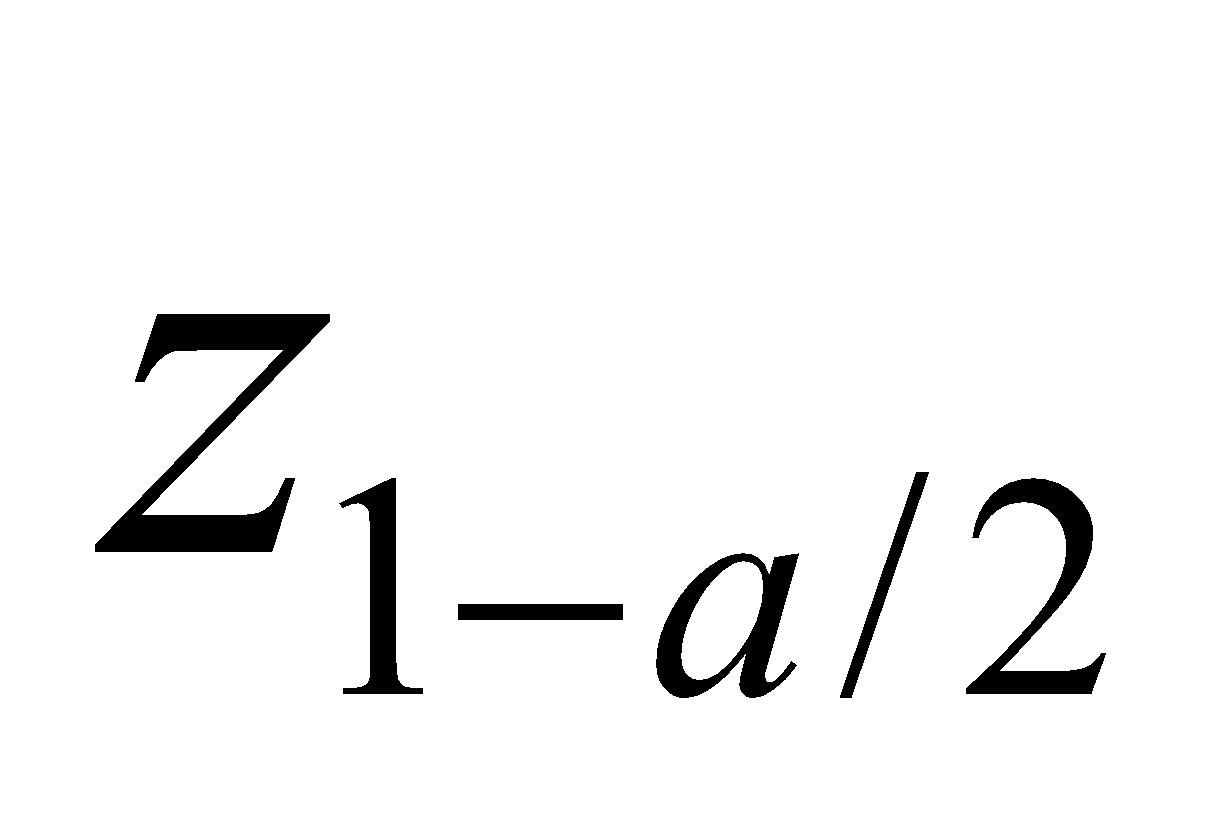
Miền bác bỏ:

* Với  bác bỏ  nếu  hoặc .
* Với  bác bỏ nếu .
* Với  bác bỏ nếu .

Trong R, để tìm phân vị  sử dụng hàm qt(1-alpha/2,n-1).

Trong kết quả do R xuất ra, ta xác định có bác bỏ  hay không thông qua P- giá trị.

Quy tắc: Khi P- giá trị bé hơn thì bác bỏ .

Khi cỡ mẫu lớn, phân phối của thống kê  sẽ xấp xỉ phân phối chuẩn hóa N(0,1), khi đó giá trị tiêu chuẩn dùng để so sánh là (dùng qnorm(1-alpha/2)).

Sử dụng hàm t.test để kiểm định theo cú pháp



*Hình 1.4: Cú pháp kiểm định cho kỳ vọng một mẫu*

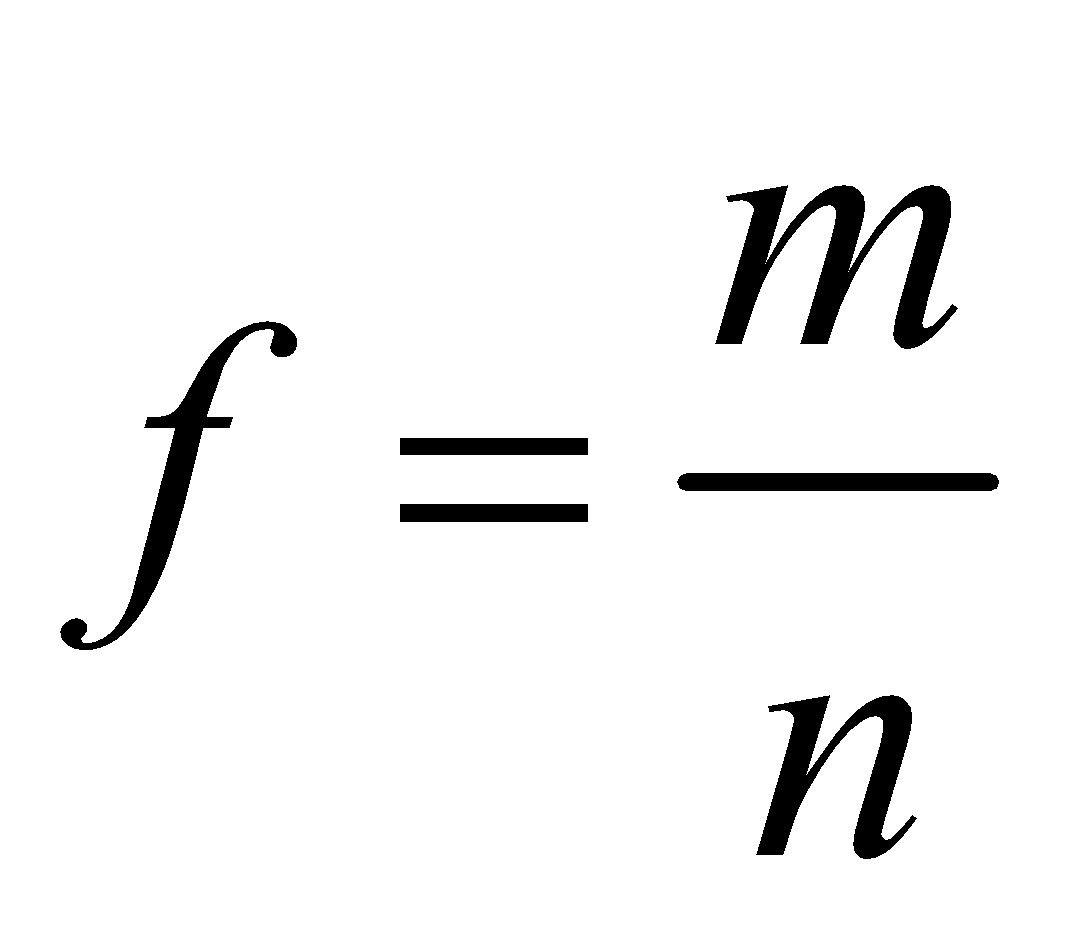
o x: véc tơ dữ liệu.

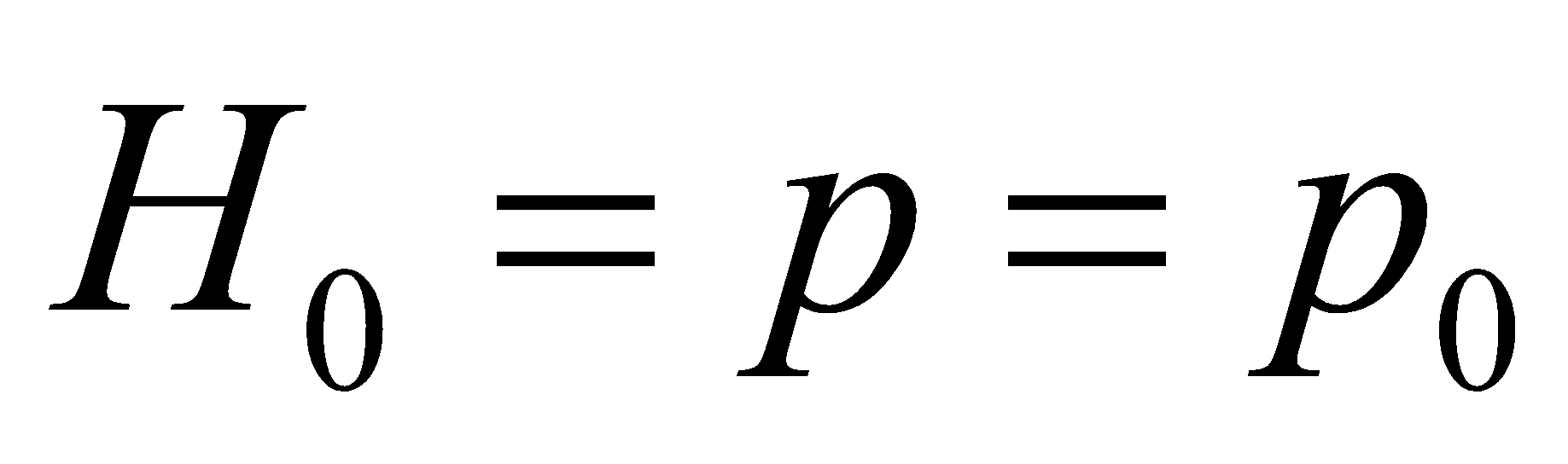
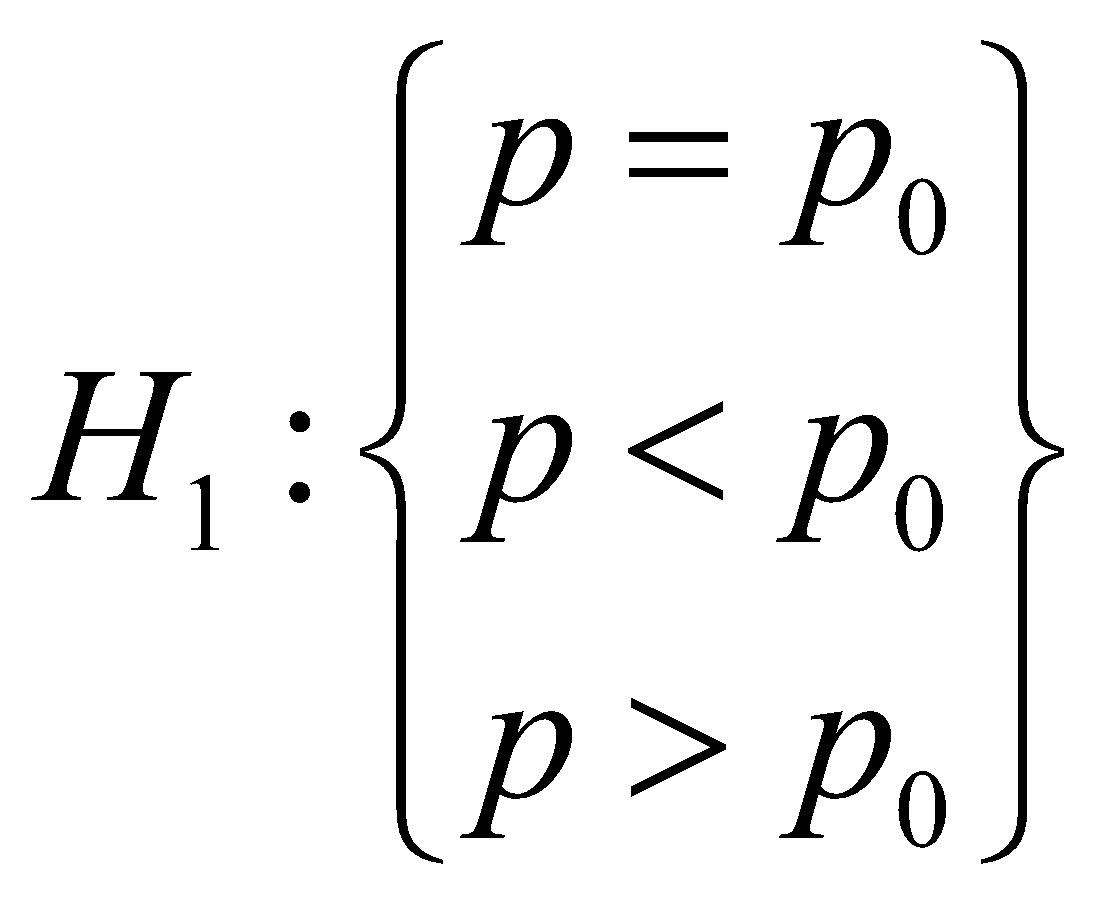
o alternative: xác đӏnh kiểm định là hai phía (“two.sided”), bên trái (“less”) hay bên phải (“greater”), mặc định là two.sided.

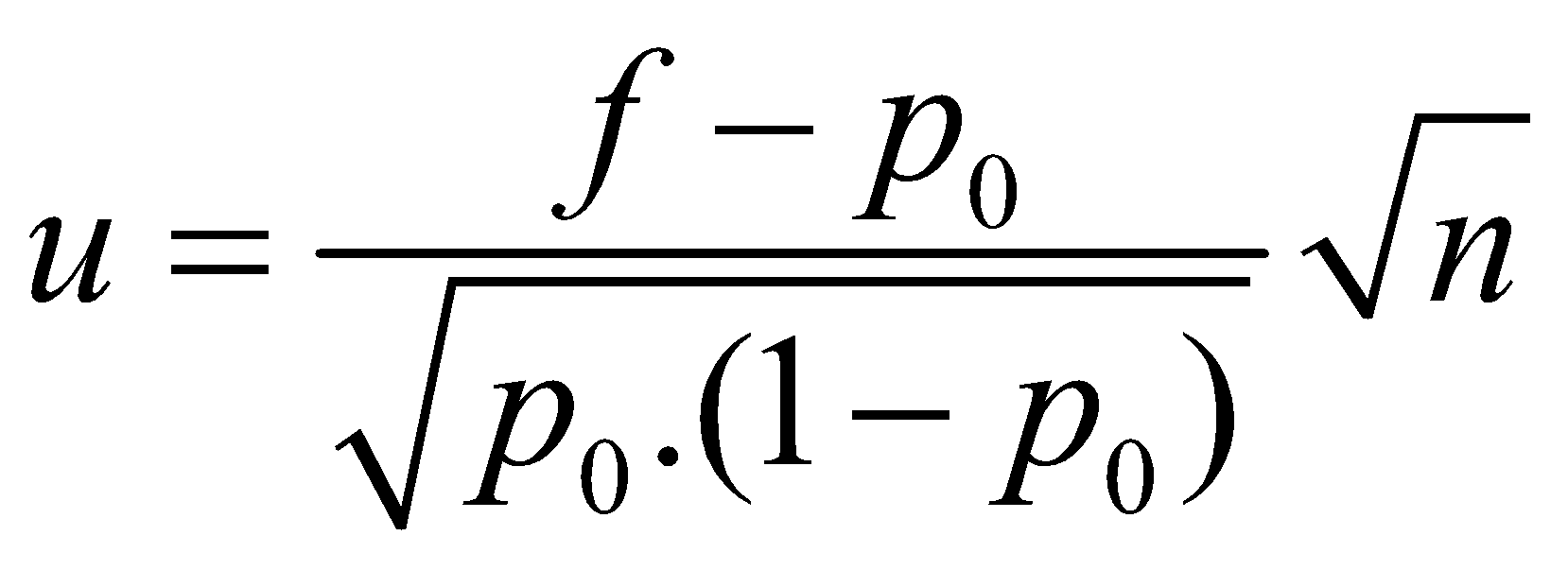
o mu = mu\_0: giá trị cần kiểm định.

o conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.

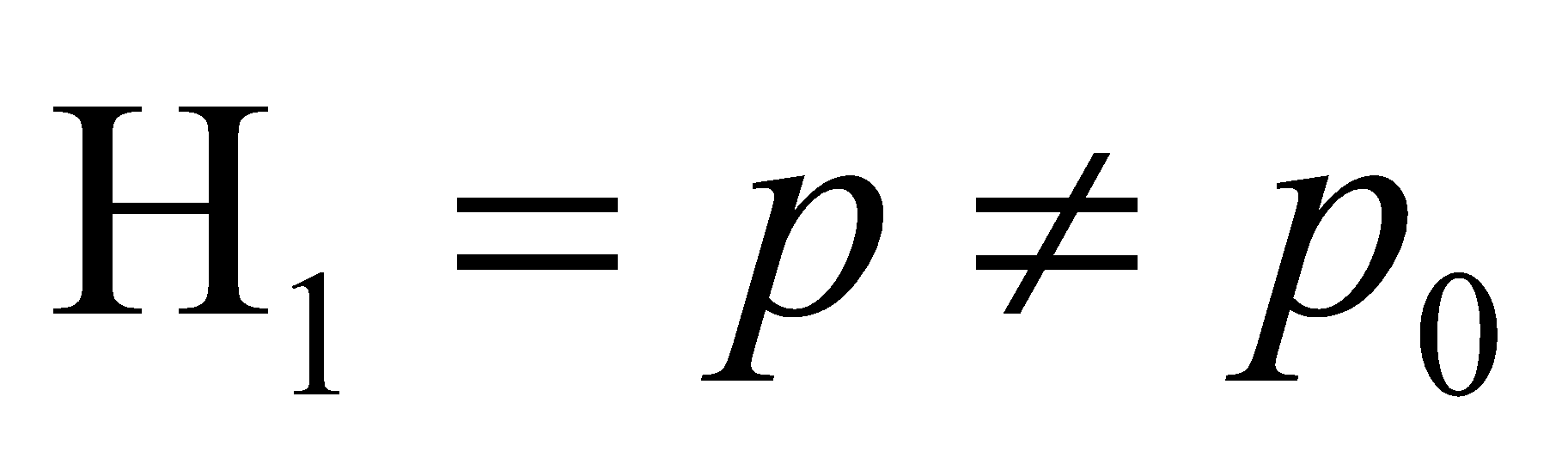
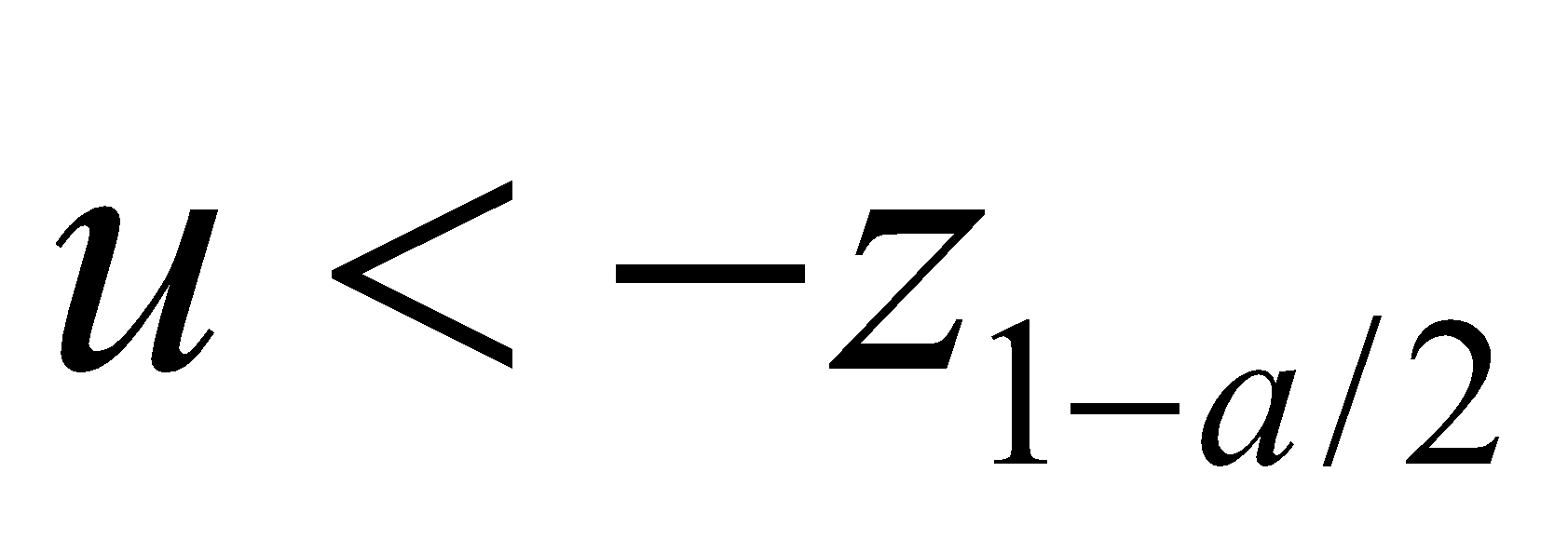
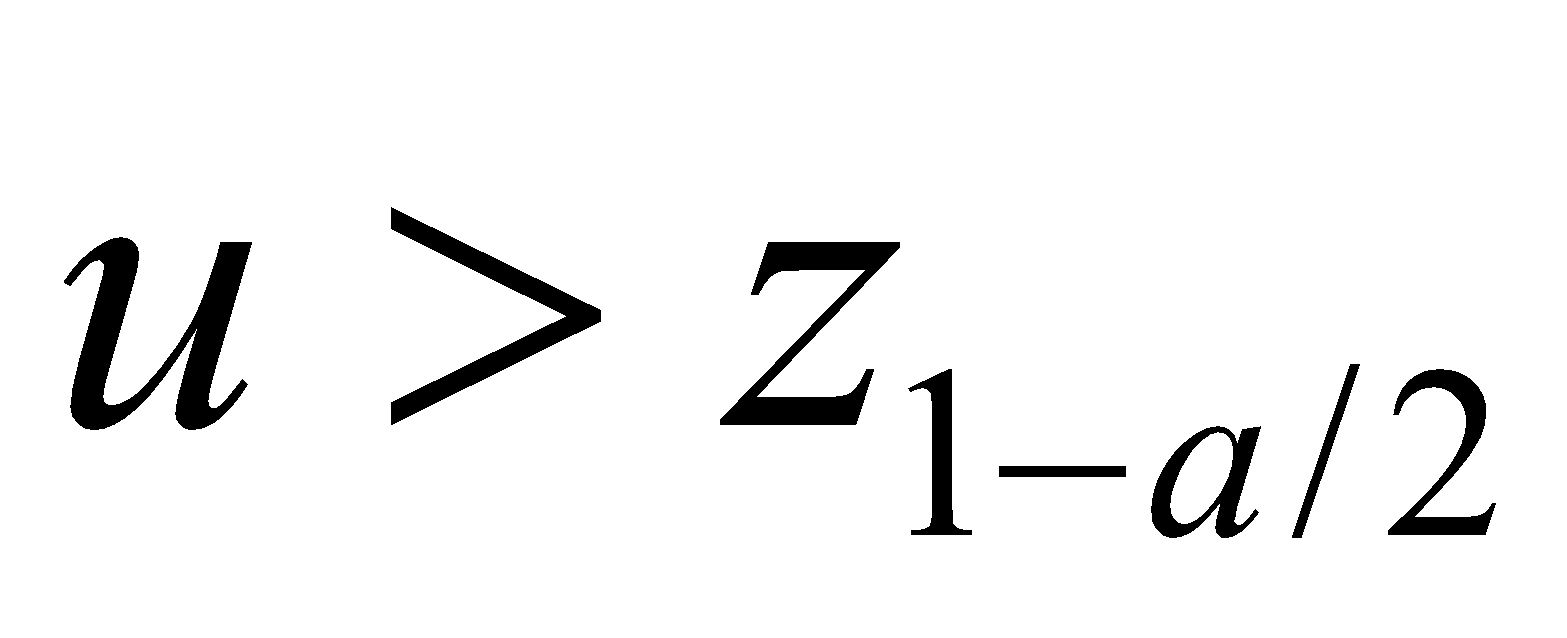
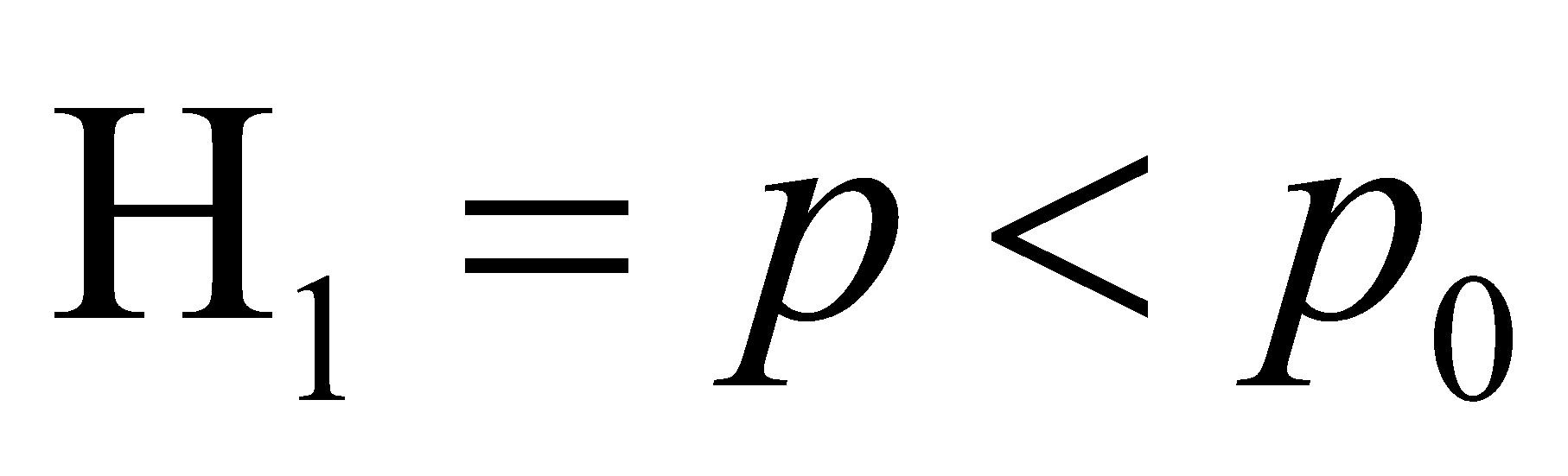
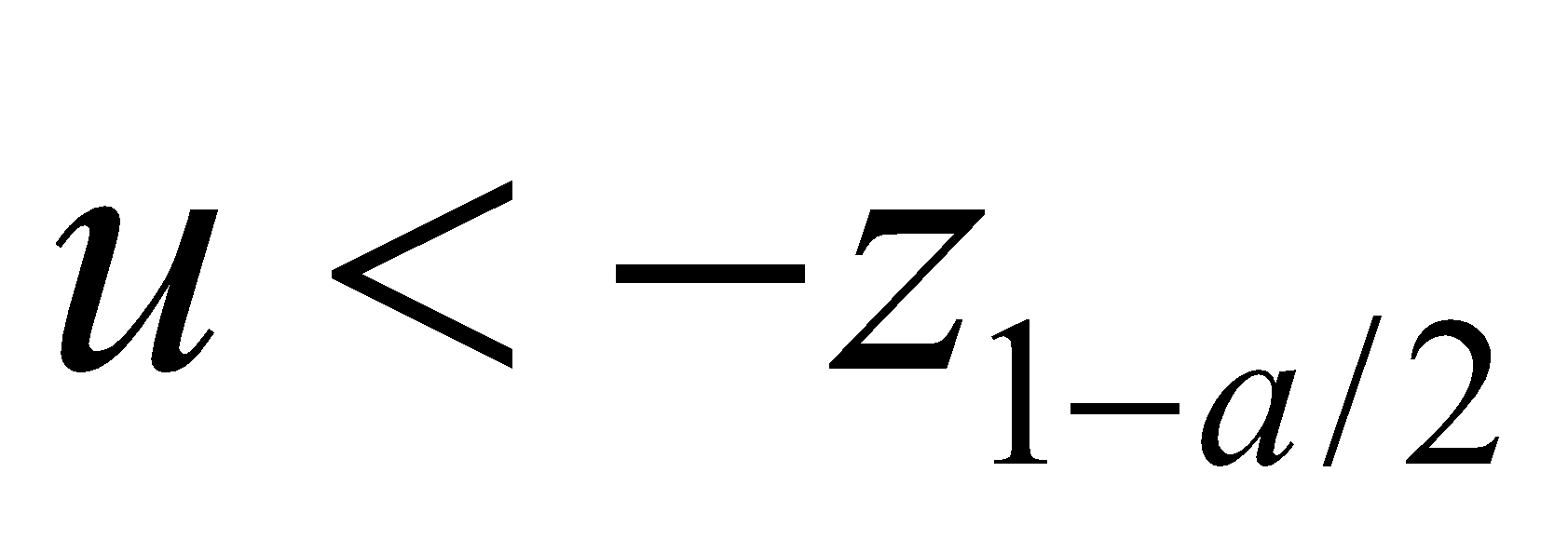
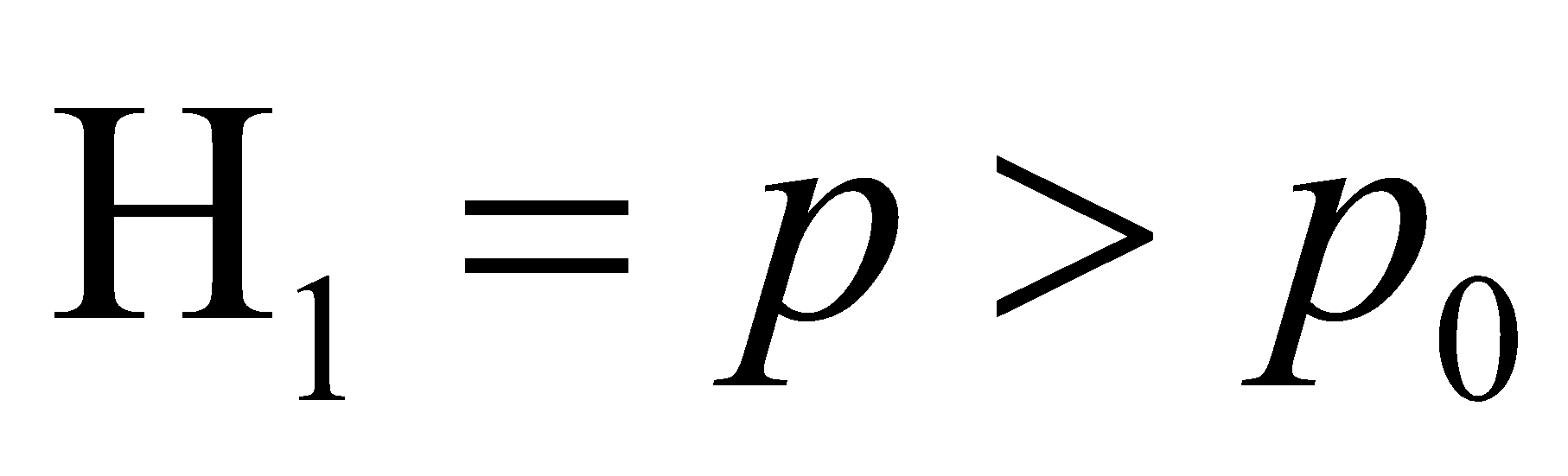
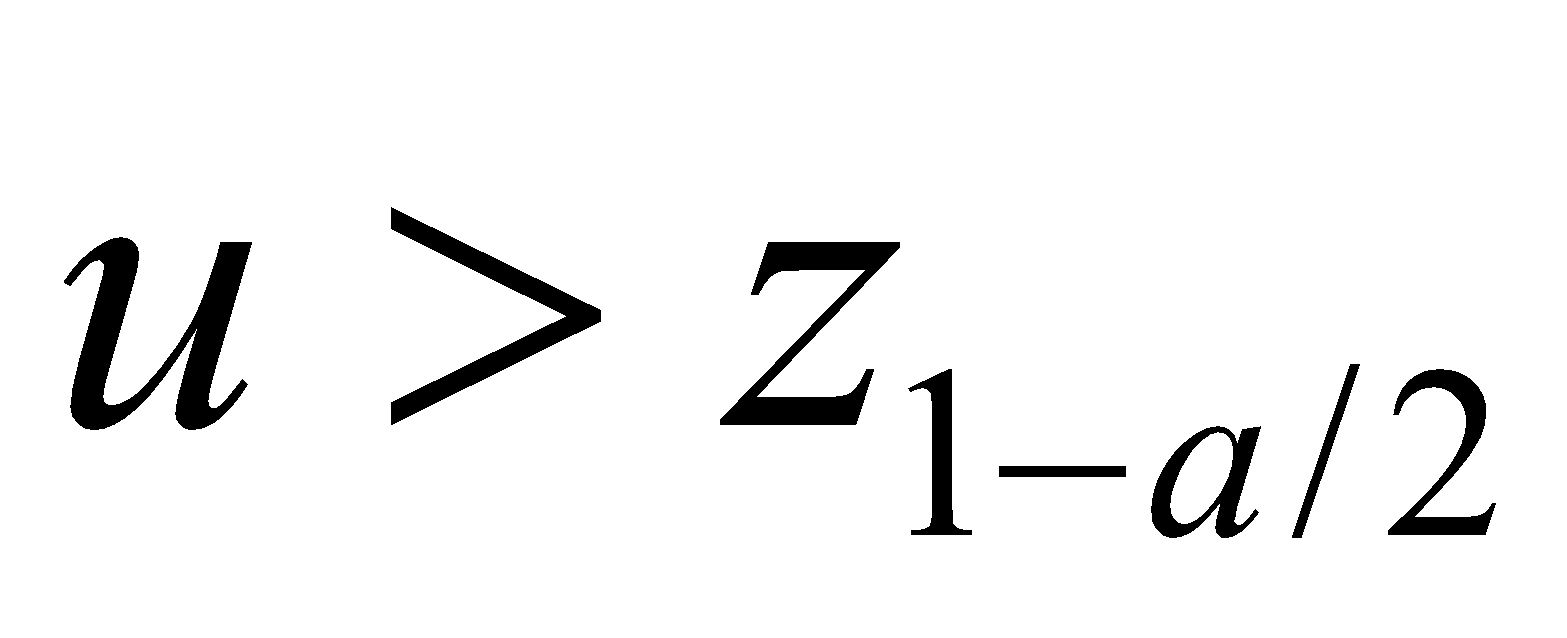
* **Kiểm định giả thuyết cho tỷ lệ một mẫu**

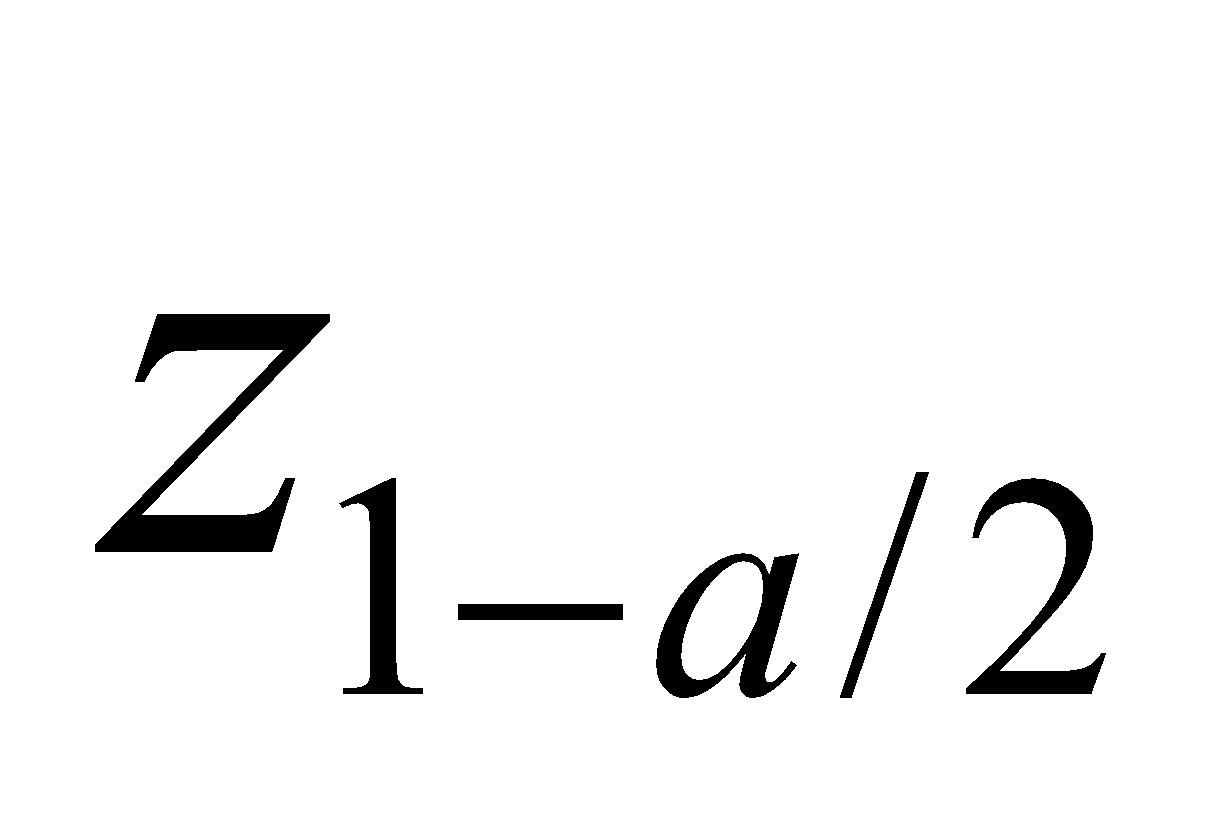
Giả sử cần kiểm định tỷ lệ phần tự thỏa mãn tính chất A trong tập thể. Khảo sát một số mẫu n. Gọi m là tổng số phần tử thỏa mãn tính chất A trong n phần tử khảo sát, suy ra tỷ lệ mẫu:  Giả thuyết cỡ mẫu khảo sát n phải tương đối lớn.

Giả thuyết: ; Đối thuyết: (Một trong 3 trường hợp).

Tính thống kê kiểm định: .

Miền bác bỏ:

* Với  bác bỏ  nếu  hoặc .
* Với  bác bỏ  nếu .
* Với  bác bỏ  nếu .

Để tìm , sử dụng hàm qnorm(1-alpha/2).

Sử dụng hàm prop.test để kiểm định:



*Hình 1.5: Cú pháp kiểm định giả thuyết cho tỷ lệ một mẫu*

Trong đó:

o m: số phần tử thỏa mãn tính chất A trong n phần tử của mẫu.

o n: kích thước mẫu.

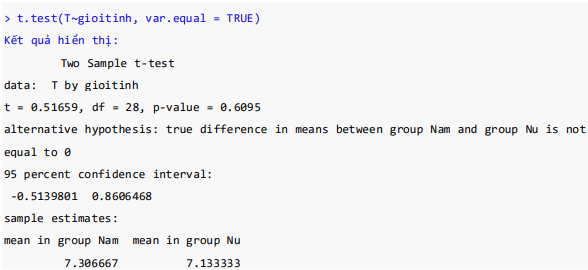
o alternative: xác định kiểm định là hai phía (“two.sided”), bên trái (“less”) hoặc bên phải (“greater”).

o p = p0: giá trị cần kiểm định.

o conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.

* **Kiểm định trung bình hai mẫu**

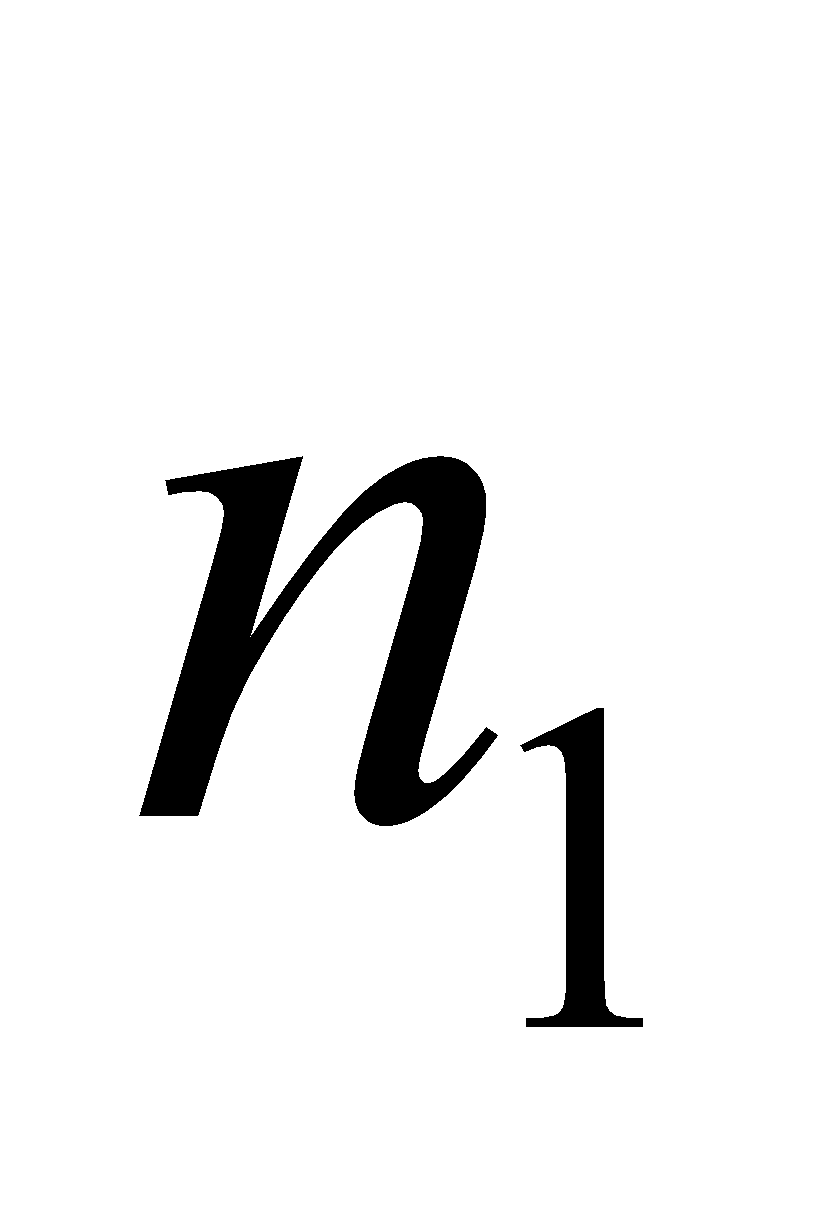
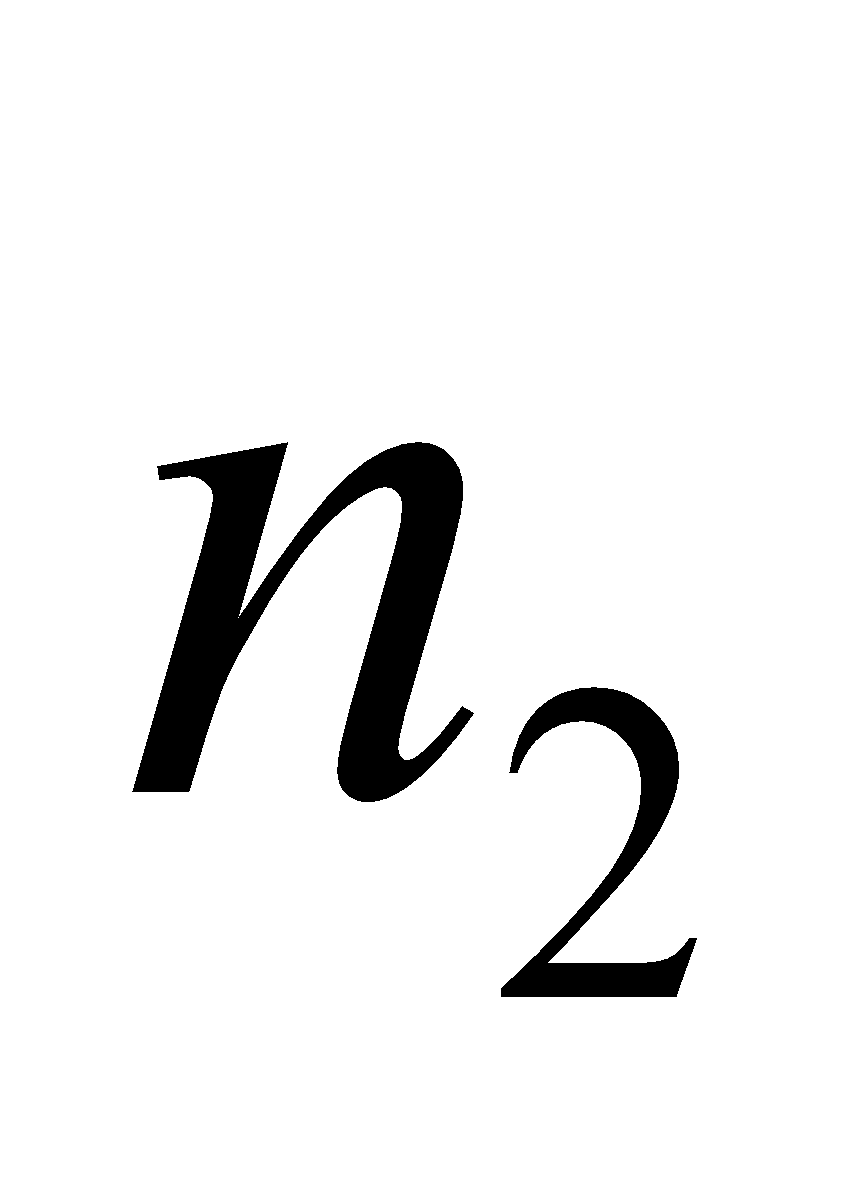
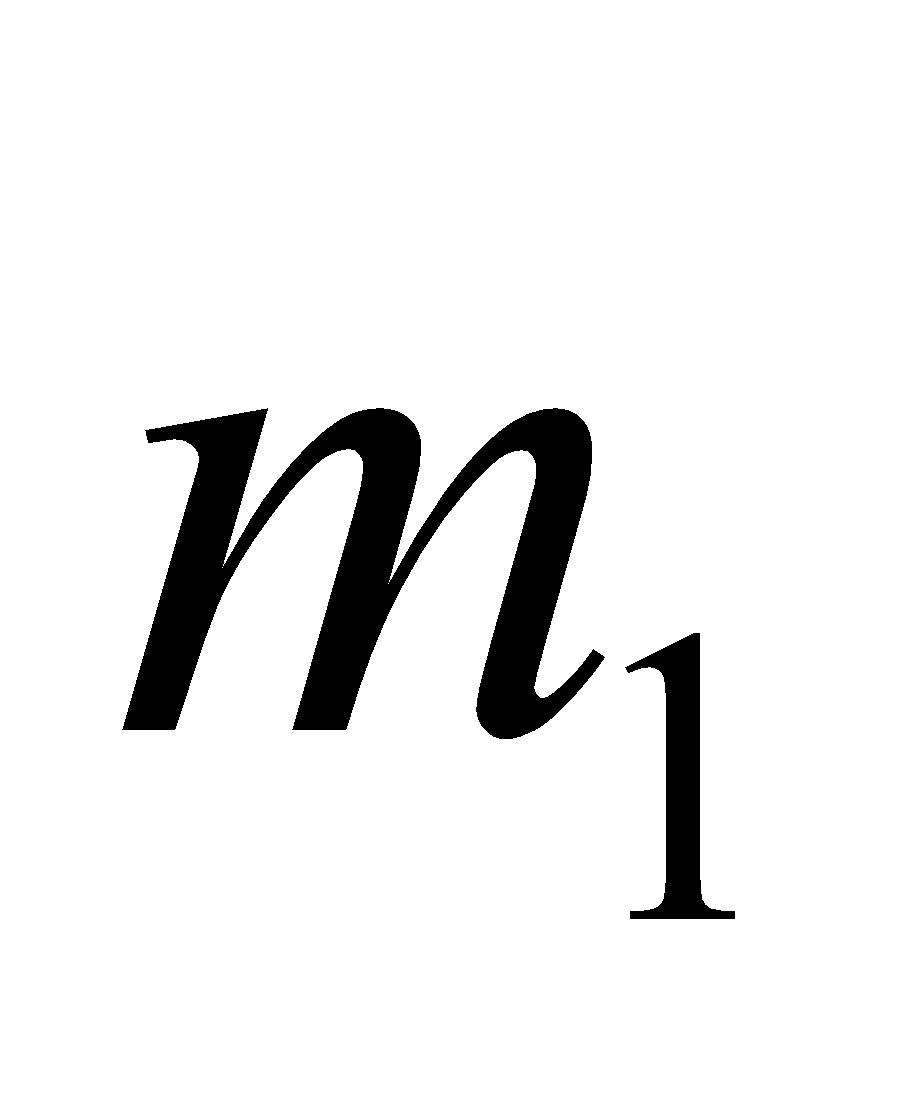
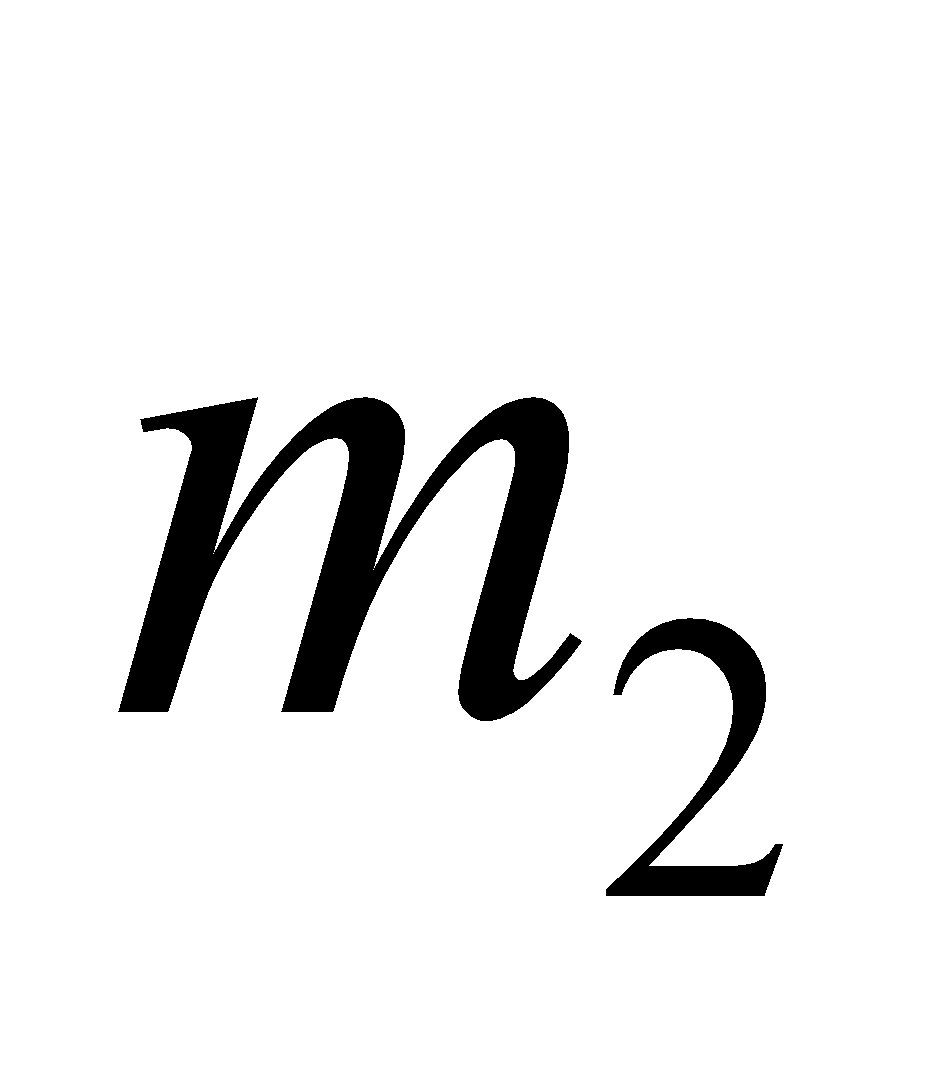
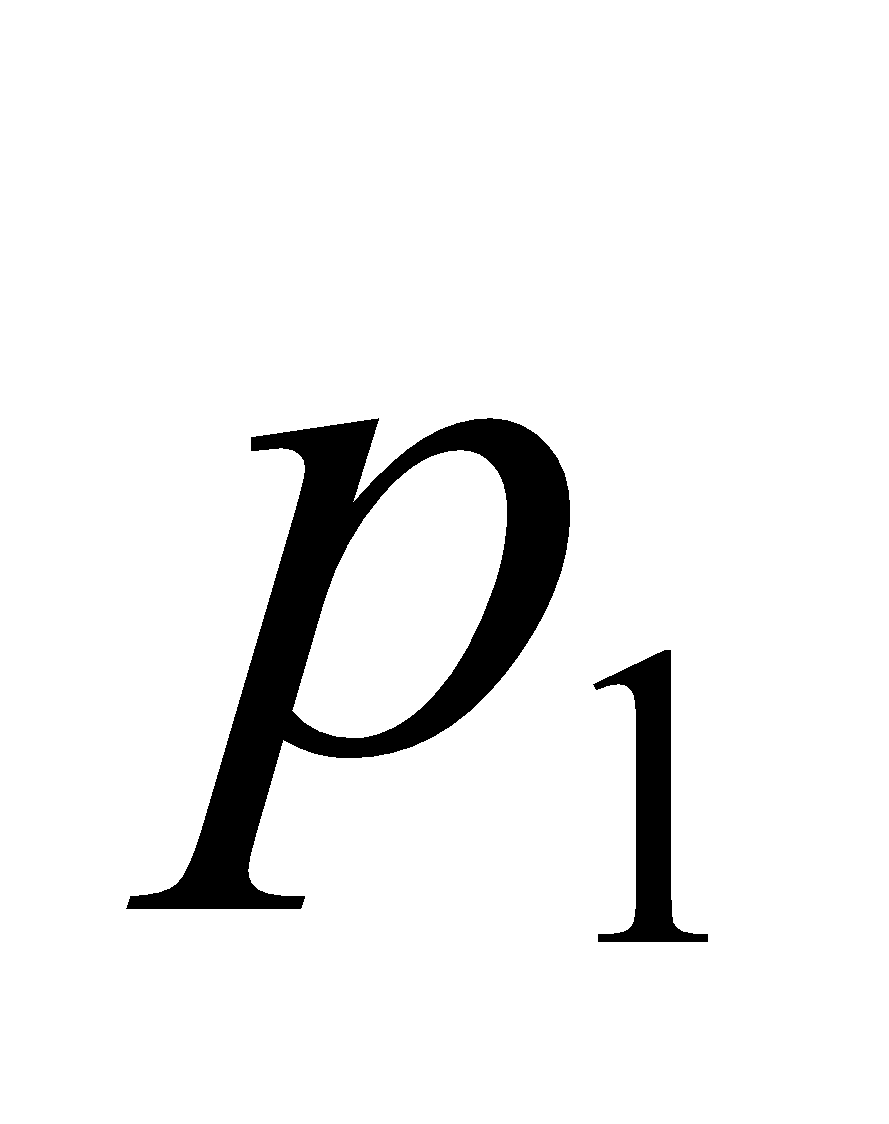
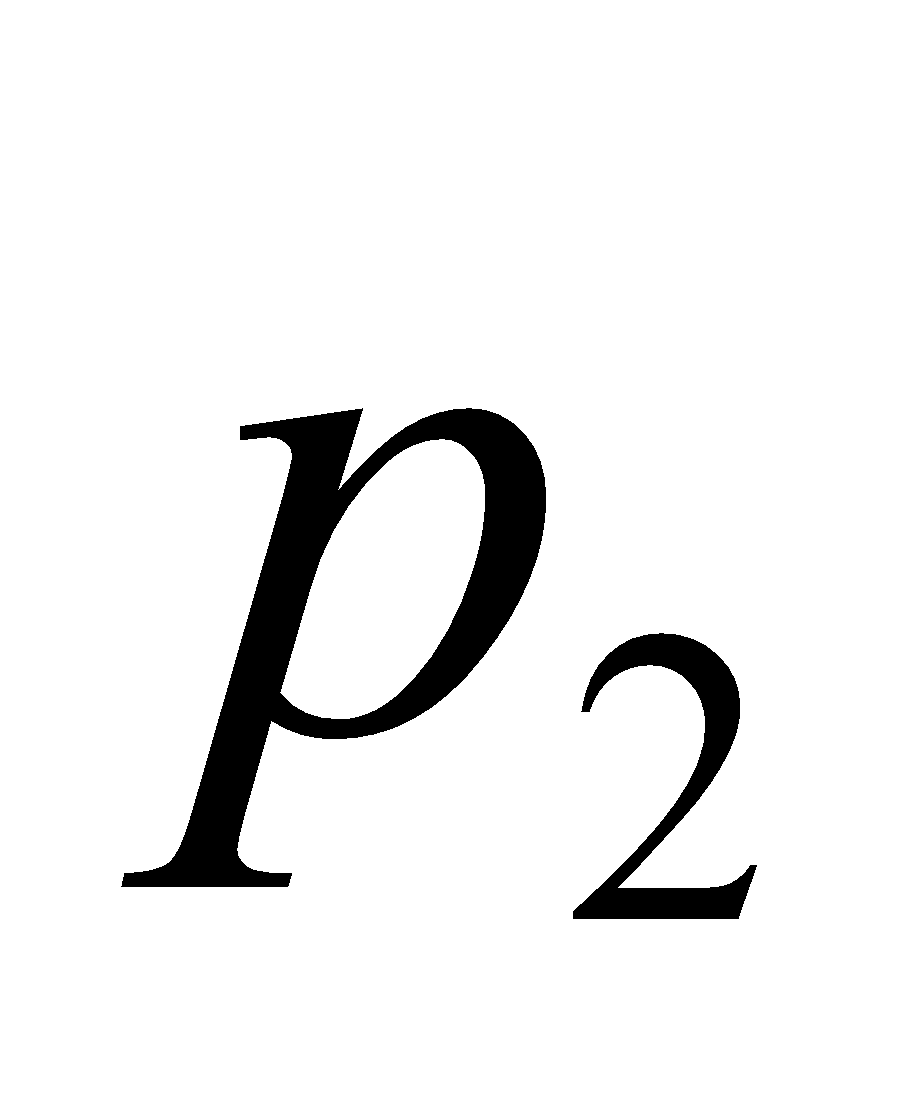
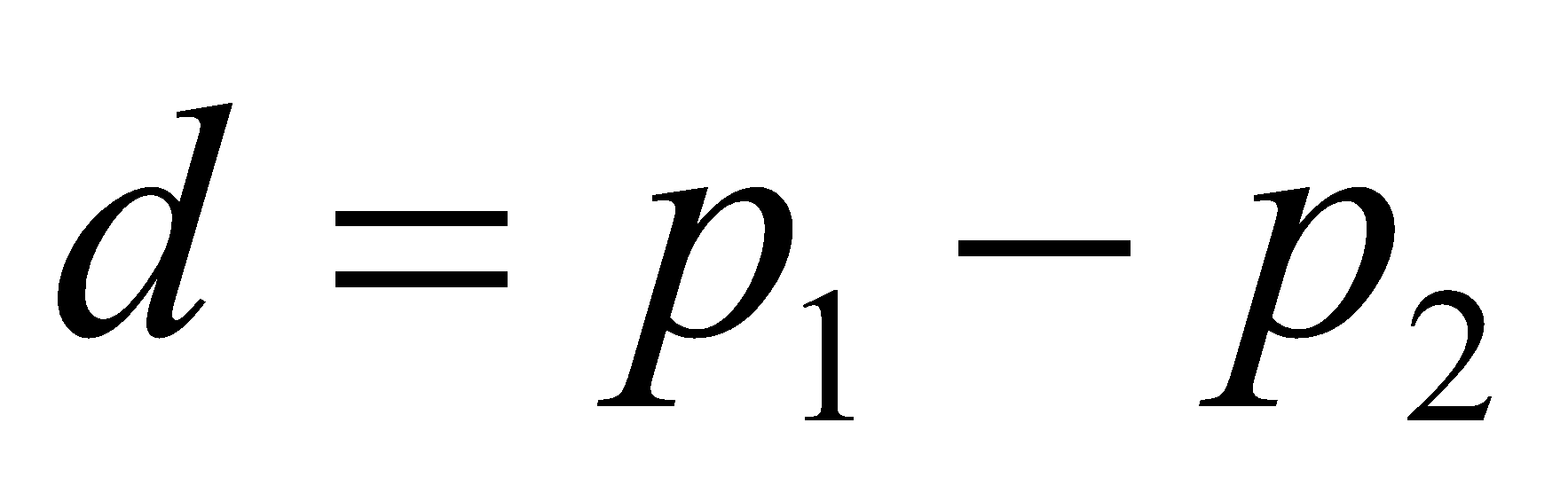
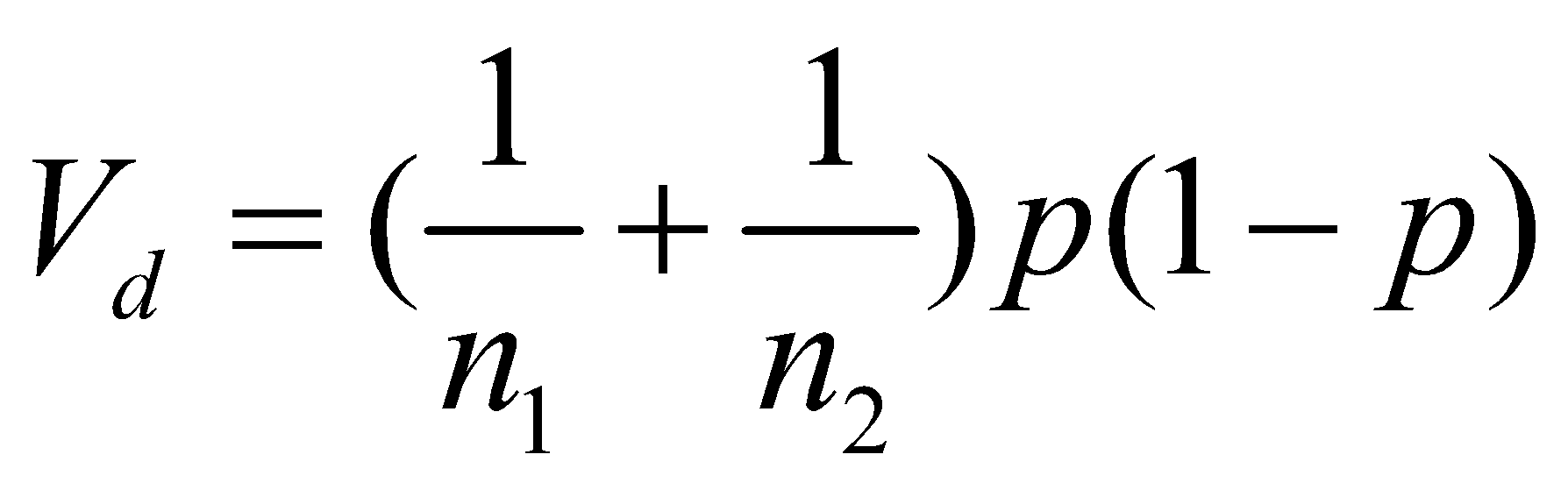
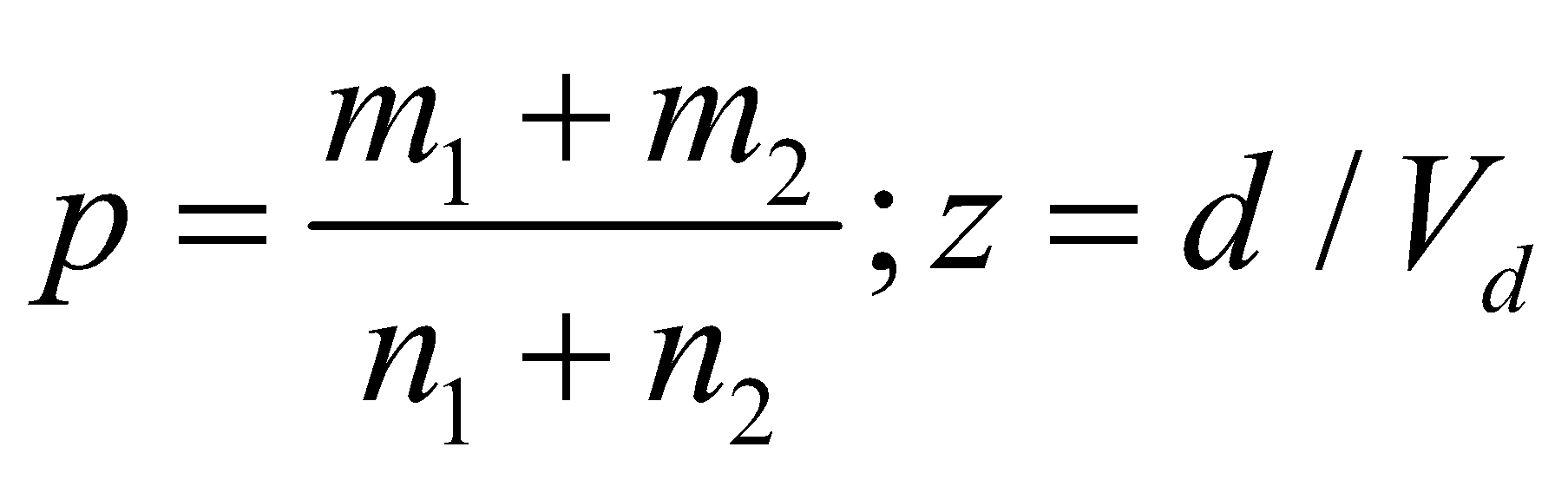
Kiểm định trên dựa vào giả định hai nhóm nam và nữ có phương sai. Nếu chúng ta có lý do để tin rằng hai nhóm có cùng phương sai, chúng ta chỉ có thể thay đổi một tham số trong hàm t với var.equal = TRUE như sau:



*Hình 1.6: Kết quả của trung bình hai mẫu*

Về mặt số liệu, kết quả phân tích trên có khác chút ít so với kết quả phân tích dựa vào giả định hai phương sai khác nhau, nhưng trị số p cũng đưa đến một kết luận rằng sự khác biệt giữa hai nhóm không có ý nghĩa thống kê.

* **Kiểm định tỷ lệ hai mẫu**

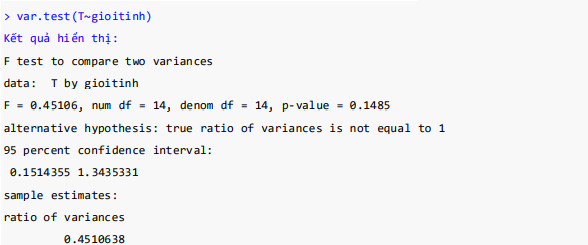
Cho hai mẫu với số đối tượng  và  , gọi số phần tử thỏa mãn tính chất A trong mẫu 1 là  ,trong mẫu 2 là  .Do đó, chúng ta có thể tính được tỉ lệ tương ứng trong hai mẫu là , .Lý thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu  tuân theo luật phân phối chuẩn với số trung bình 0 và phương sai bằng: .Trong đó:  tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1.

*d) Kiểm định Wilcoxon cho hai mẫu (wilcox.test)*

Kiểm định t dựa vào giả thiết là phân phối của một biến phải tuân theo luật phân phối chuẩn. Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lý.

*e) So sánh phương sai (var.test)*

***Ví dụ:*** Sử dụng tập tin dữ liệu "Diem\_TN", để kiểm định phương sai điểm toán (T) giữa hai nhóm nam và nữ có khác nhau không, chúng ta sử dụng câu lệnh sau:



*Hình 1.7 : Kết quả của so sánh phương sai*

Kết quả trên cho thấy độ khác biệt về phương sai giữa hai nhóm là 0.45 lần. Trị số p = 0.1485 cho thấy phương sai giữa hai nhóm khác nhau không có ý nghĩa thống kê.

*f) Thủ tục kiểm định shapiro.test về phân phối chuẩn*

Để kiểm định một luật phân phối mẫu xem liệu có tuân theo luật chuẩn hay không, chúng ta có thể sử dụng hàm shapiro.test có cấu trúc như sau:



*Hình 1.8: Thủ tục kiểm định*

### **1.2.3. Thống kê suy diễn trong các bài toán phân tích tương quan**

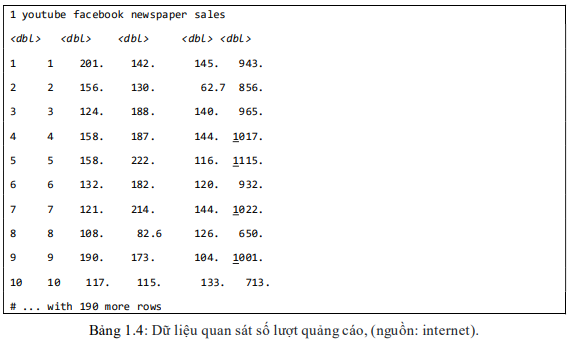
Hệ số tương quan (r) là một chỉ số thống kê đo lường mức độ liên hệ tương quan giữa hai biến số. Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hoặc gần 0), có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại, nếu hệ số tương quan bằng -1 hoặc 1 thì có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm (r < 0), có nghĩa là hai biến tương quan nghịch (khi một biến tăng thì biến kia giảm và ngược lại); nếu giá trị của hệ số tương quan là dương (r > 0), có nghĩa là hai biến tương quan thuận (khi một biến tăng thì biến kia cũng tăng, và khi một biến giảm thì biến kia cũng giảm).

Có nhiều hệ số tương quan trong thống kê, nhưng ở đây chúng ta sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson (r), Spearman (p), và Kendall (t). Trong tiểu mục này, dữ liệu dùng để minh họa là file dữ liệu "marketing.csv" có thể được tham khảo từ link:

<https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu>



*Hình 1.9: Đọc file marketing.csv*



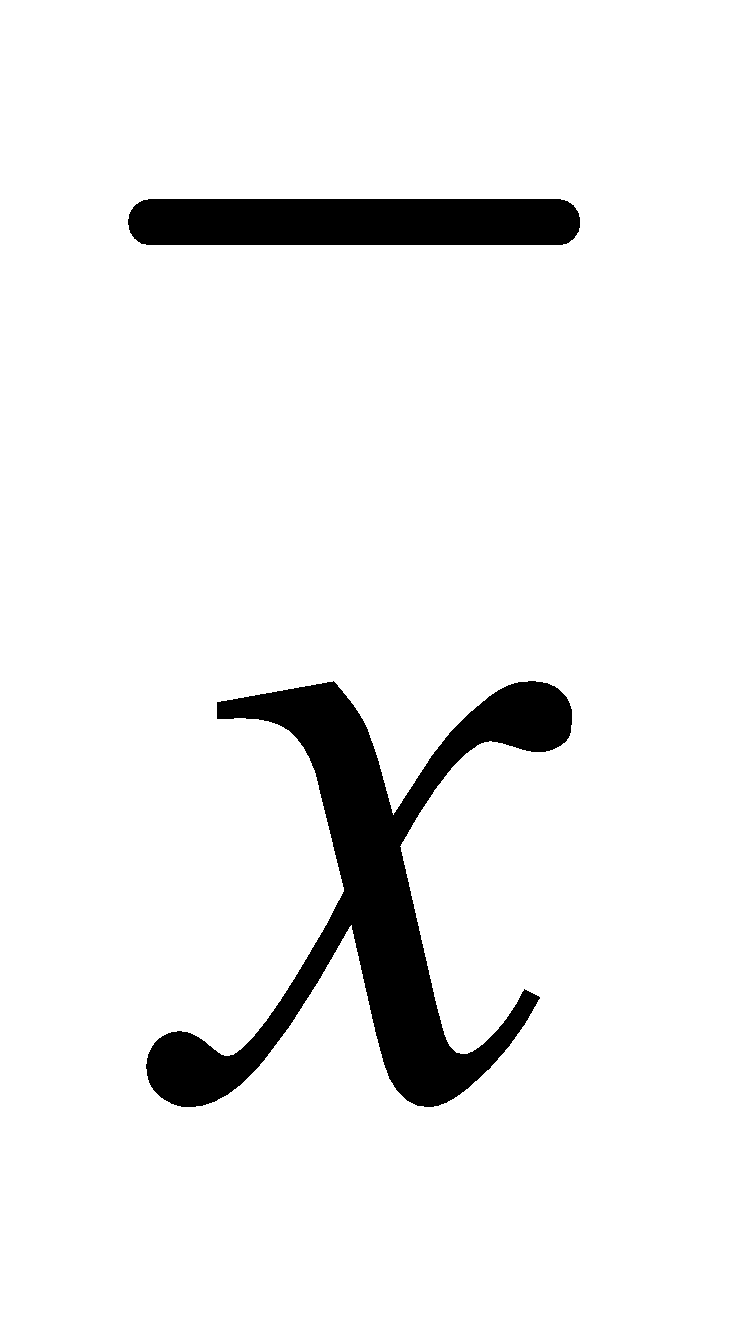
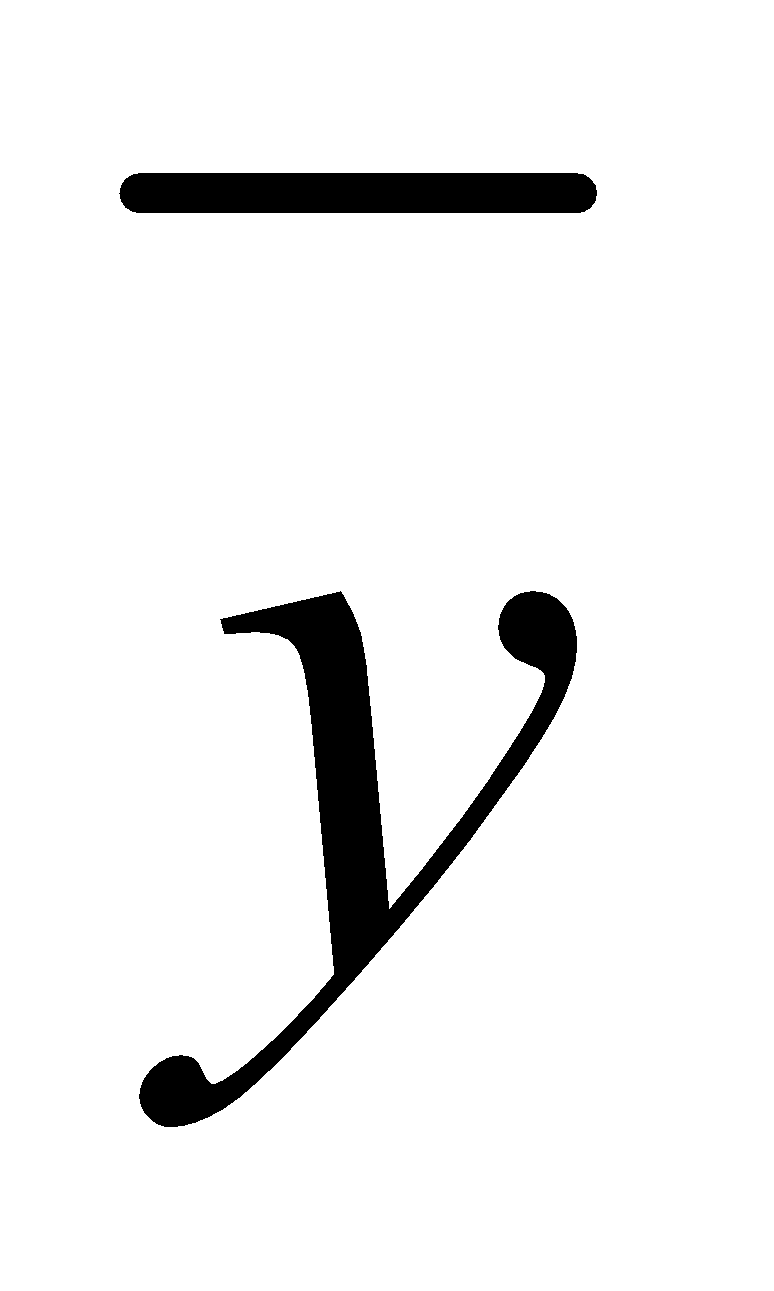
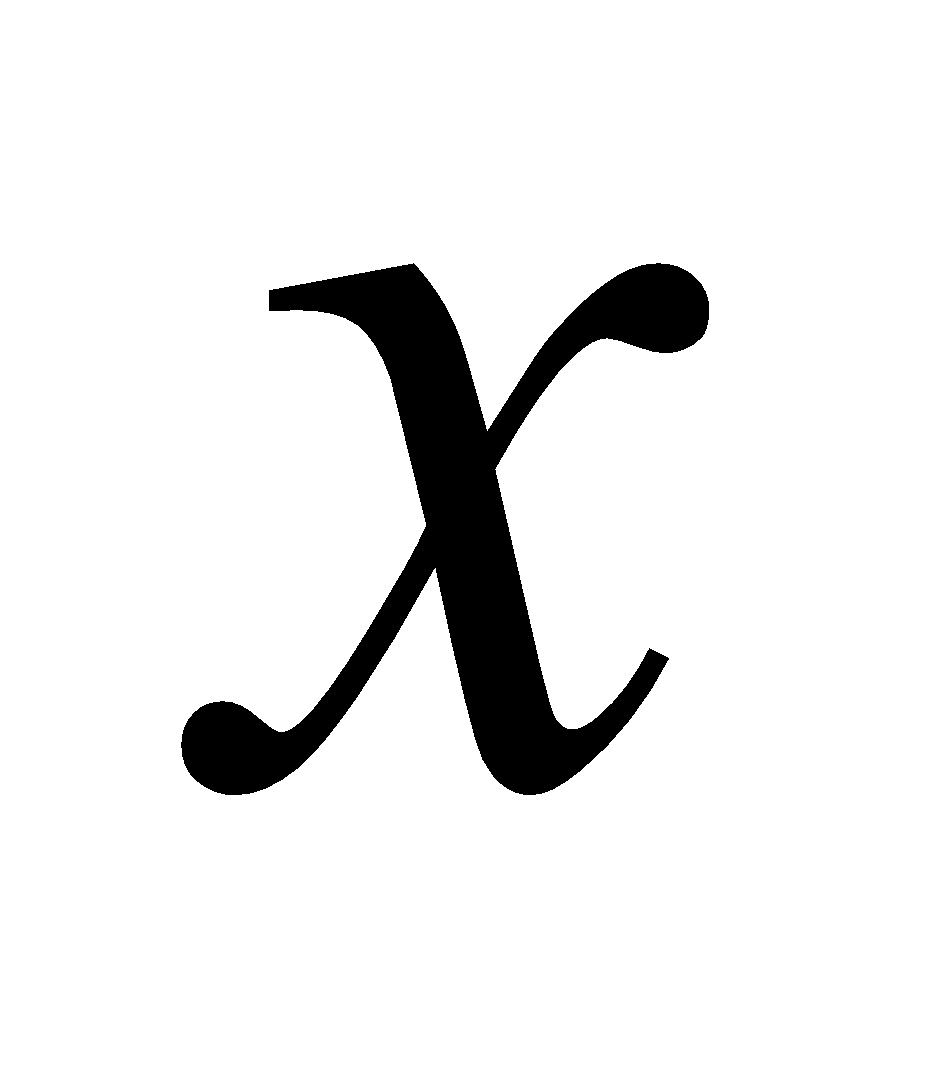
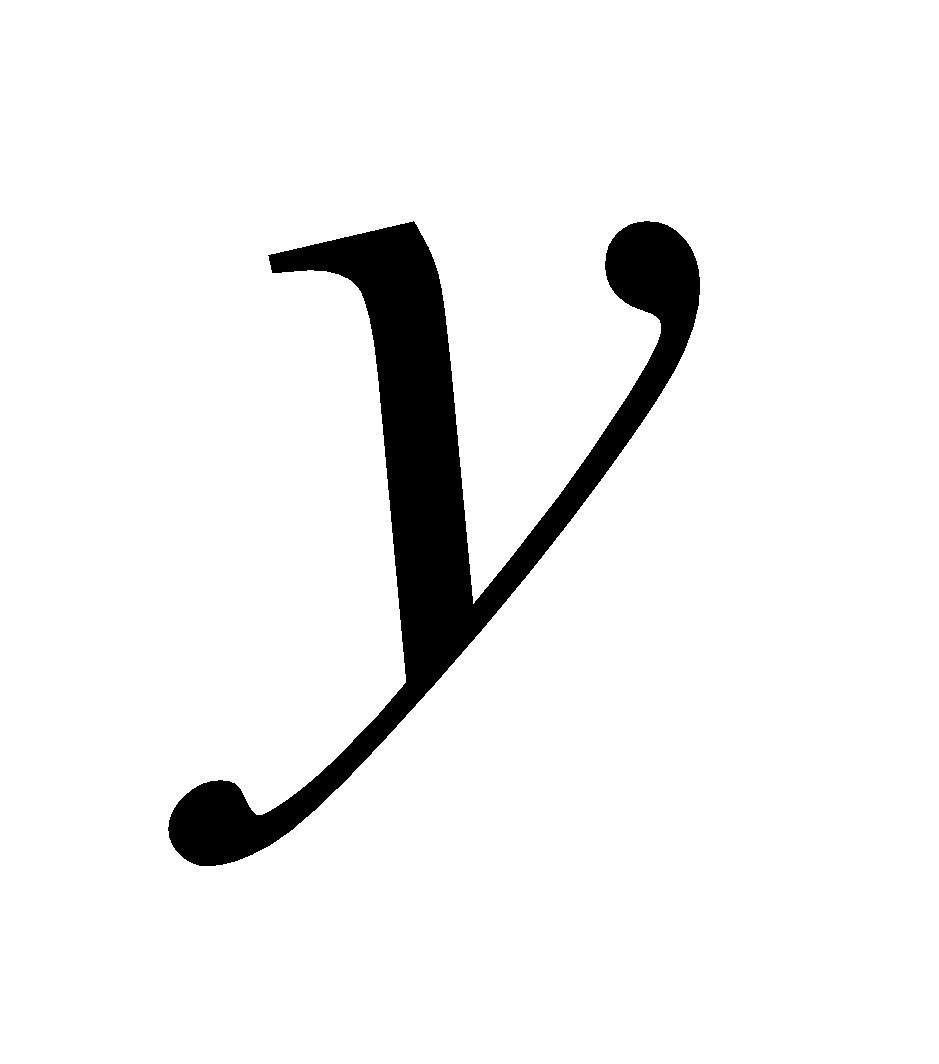
*Bảng 1.4: Dữ liệu quan sát số lượt quảng cáo, (nguồn: internet)*

1. *Hệ số tương quan mẫu*

* **Hệ số tương quan Pearson**

Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được tính bằng công thức sau đây:



Trong đó,  và  là giá trị trung bình của biến số  và 

Để tính hệ số tương quan Pearson trong R, cú pháp như sau:



*Hình 1.10: Cú pháp tính hệ số tương quan Pearson trong R*

* **Hệ số tương quan Spearman ( p )**

Hệ số tương quan Pearson chỉ hợp lý nếu biến x và y tuân theo luật phân phối chuẩn. Nếu x và y không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác được gọi là Spearman, một phương pháp phân tích phi tham số. Hệ số này được tính bằng cách biến đổi hai biến x và y thành thứ bậc (rank), và xem xét tương quan giữa hai dãy thứ bậc. Do đó, hệ số này còn có tên tiếng Anh là Spearman’s Rank correlation.

Để tính hệ số tương quan Spearman trong R, cú pháp như sau:



*Hình 1.11: Cú pháp tính hệ số tương quan Spearman trong R*

* **Hệ số tương quan Kendall (t)**

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được tính bằng cách tìm các cặp ( x, y) "song hành" với nhau. Một cặp (x, y) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trục hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trục tung. Nếu hai biến x và y không có liên hệ với nhau, thì cặp này bằng hoặc tương đương với cặp không song hành.

Vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian tính toán của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng.

Để tính hệ số tương quan Kendall trong R, cú pháp như sau:



*Hình 1.12: Cú pháp tính hệ số tương quan Kendall trong R*

1. *Kiểm định hệ số tương quan*

Bên cạnh việc tính các giá trị tương quan mẫu, chúng ta cũng có thể kiểm định hệ số tương quan lý thuyết với giả thuyết kiểm định:

* H0: Không có tương quan ( hệ số tương quan = 0 ).
* H1: Có tương quan.

Để tính kiểm định trong R, có thể sử dụng cú pháp sau:



*Hình 1.13: Cú pháp tính kiểm định trong R*

Trong đó:

* Nhân tố 1, nhân tố 2 là 2 biến cần kiểm định tính tương quan.
* methol được lựa chọn một trong ba phương pháp tương ứng.

# **CHƯƠNG 2: PHÂN TÍCH THĂM DÒ DỮ LIỆU**

## **2.1. GIỚI THIỆU BÀI TOÁN**

### **2.1.1. Giới thiệu bài toán**

Trong bối cảnh ngành công nghiệp thực phẩm nhanh đang phát triển mạnh mẽ, McDonald's nổi lên như một trong những chuỗi cửa hàng ăn uống lớn nhất và có tầm ảnh hưởng nhất trên thế giới. Với sự hiện diện toàn cầu, McDonald's không chỉ là một biểu tượng của thức ăn nhanh mà còn là một kho dữ liệu khổng lồ về thực phẩm, dinh dưỡng và xu hướng tiêu dùng.

Bài toán này tập trung vào việc phân tích và khám phá dữ liệu dinh dưỡng từ thực đơn của McDonald's. Mục tiêu là hiểu rõ hơn về giá trị dinh dưỡng của các sản phẩm, so sánh giữa các nhóm thực phẩm khác nhau và tìm ra những xu hướng thú vị có thể ảnh hưởng đến quyết định kinh doanh cũng như lựa chọn của người tiêu dùng.

Bằng cách sử dụng các phương pháp thống kê mô tả và trực quan hóa dữ liệu, chúng ta sẽ phân tích các yếu tố như lượng calo, chất béo, protein, carbohydrate và các thành phần dinh dưỡng khác. Qua đó, chúng ta không chỉ cung cấp một cái nhìn sâu sắc về dinh dưỡng mà còn giúp đưa ra những gợi ý hữu ích cho việc cải thiện thực đơn và chiến lược tiếp thị của McDonald's.

### **2.1.2. Thu thập dữ liệu**

* Tên dữ liệu: “ India\_Menu.csv”
* Nguồn thu thập dữ liệu:

“https//://www.kaggle.com/datasets/deepcontractor/mcdonalds-india-menu-nutrition-facts"

* Kỹ thuật thu thập: tìm kiếm keywords + ngôn ngữ lập trình + keywords tiếng anh.

### **2.1.3. Phạm vi và giới hạn**

* Chỉ tập trung vào dữ liệu dinh dưỡng từ thực đơn của McDonald’s, không bao gồm dữ liệu từ các chuỗi cửa hàng khác.
* Các phân tích sẽ dựa trên thông tin hiện có, với những giới hạn nhất định về độ chính xác và tính đầy đủ của dữ liệu.

### **2.1.4. Phương pháp và công cụ**

* Sử dụng các phương pháp thống kê mô tả và trực quan hóa dữ liệu để phân tích dữ liệu dinh dưỡng.
* Các công cụ và phần mềm như R và ggplot2 sẽ được sử dụng để thực hiện các phân tích và tạo biểu đồ.

### **2.1.5. Giải thích các biến dữ liệu**

* **Menu Category**: Danh mục cho mỗi loại menu. Bao gồm các nhóm: Regular, Breakfast, McCafe, Desserts, Gourmet, Beverages và Condiments Menu.
* **Menu Items**: Các món ăn trong menu có thể tiêu thụ.
* **Per Serve Size**: Lượng tiêu chuẩn của mỗi phần dịch vụ, được đo bằng gram (g) hoặc mililit (mL).
* **Energy (kCal)**: Đơn vị dinh dưỡng đo năng lượng bằng kilocalories (kCal).
* **Protein (g)**: Giúp tăng cảm giác no, có lợi cho việc giảm cân.
* **Total fat (g)**: Giúp cảm thấy no và bảo vệ chống bệnh tim.
* **Sat Fat (g)**: Có thể tăng cholesterol xấu và triglycerides, tăng nguy cơ mắc bệnh tim (mặc dù không cần phải tránh hoàn toàn).
* **Trans fat (g)**: Chuyển hóa nhân tạo thành chất béo bão hòa, tăng nguy cơ mắc bệnh tim và đột quỵ bằng cách tăng cholesterol xấu và giảm cholesterol tốt.
* **Cholesterol (mg)**: Giúp xây dựng tế bào và sản xuất một số hormone nhất định, nhưng nếu ăn quá nhiều chất béo bão hòa và trans có thể gây tắc nghẽn động mạch.
* **Total carbohydrate (g)**: Nguồn năng lượng ưa thích của cơ thể và nhiên liệu cho các cơ quan quan trọng.
* **Total Sugars (g)**: Bao gồm đường tự nhiên và đường thêm vào.
* **Added Sugars (g)**: Đường được thêm vào, có thể ảnh hưởng đến cảm giác đói và no ít quan trọng hơn.
* **Sodium (mg)**: Giúp điều chỉnh cân bằng nước, nhưng nếu ăn quá nhiều có thể góp phần vào việc giữ nước.

## **2.2. XỬ LÝ DỮ LIỆU**

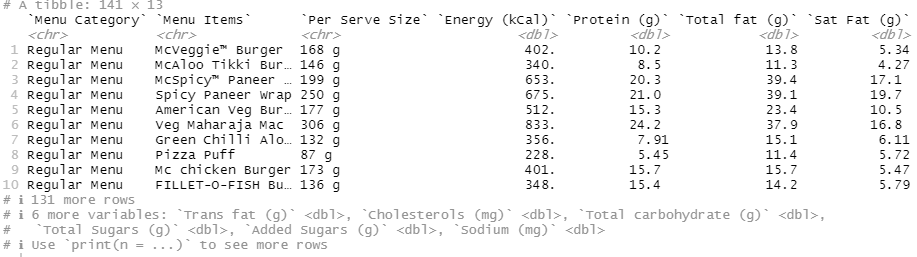
### **2.2.1. Thu thập dữ liệu**

Sử dụng hàm ‘ read\_csv ‘ để nhập dữ liệu vào môi trường phân tích:

library(readr)

data <- readr::read\_csv("D:/ĐỒ ÁN (1)/India\_Menu.csv")

print(data)



*Bảng 2.1: Dữ liệu dinh dưỡng từ thực đơn McDonald’s*

### **2.2.2. Kiểm tra dữ liệu**

Sử dụng các hàm như ‘ head() ‘ , ‘ tail() ‘ , ‘ str() ‘, hoặc ‘ summary() ‘ để hiển thị cấu trúc và các loại biến trong tập dữ liệu:

#Hiển thị 6 dòng đầu tiên

head(data)

#Hiển thị 6 dòng cuối cùng

tail(data)

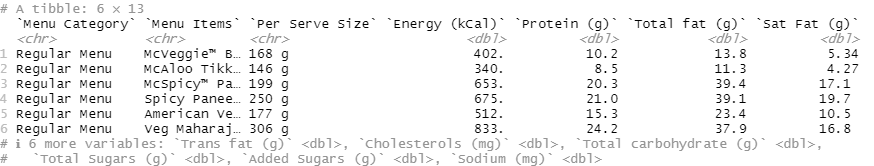
#Hiển thị cấu trúc data

str(data)

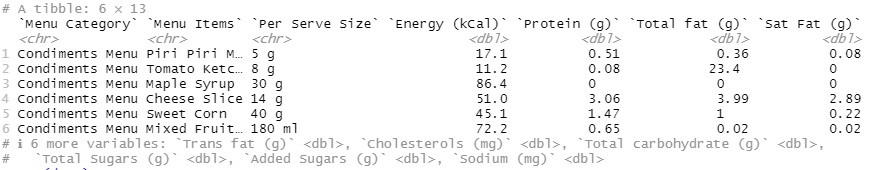
#Hiển thị tóm tắt thống kê

summary(data)

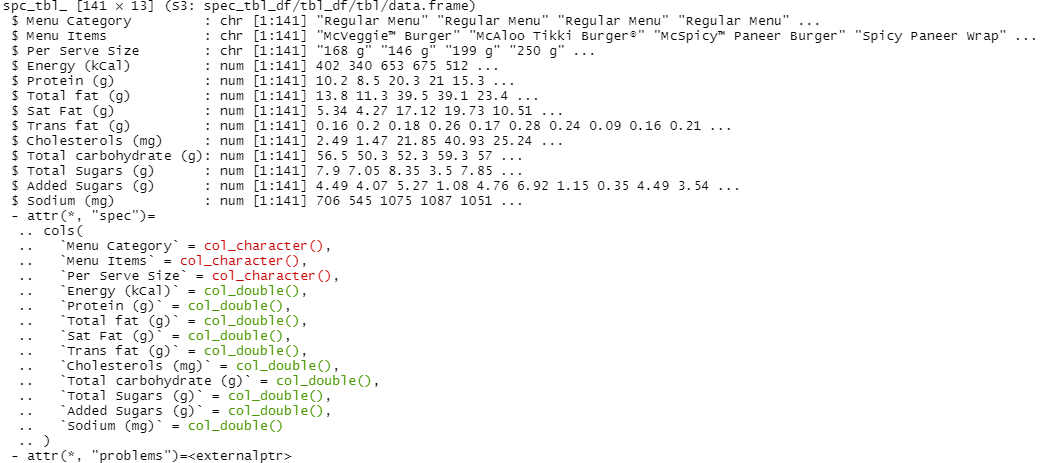
Kết quả hiển thị:



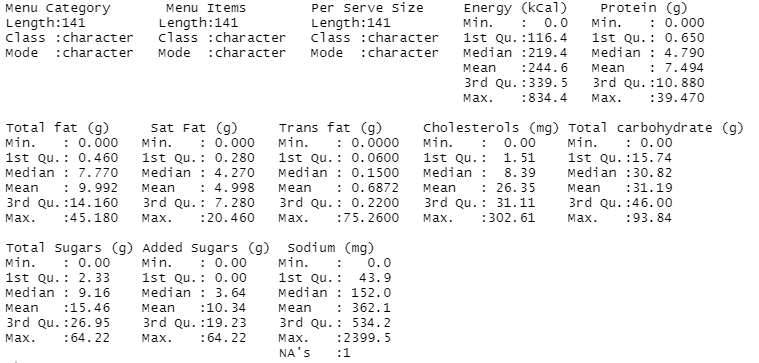
*Bảng 2.2: Hiển thị 6 dòng đầu của data*



*Bảng 2.3: Hiển thị 6 dòng cuối của data*



*Bảng 2.4: Hiển thị cấu trúc của data*



*Bảng 2.5: Tóm tắt thống kê cho từng cột trong data*

### **2.2.3. Làm sạch dữ liệu**

Nhìn vào bảng 2.5, ta có thể thấy ở cột Sodium có một giá trị NA, cho nên ta sẽ thay thế giá trị NA bằng giá trị trung bình:

# Tải thư viện dplyr

library(dplyr)

# Thay thế giá trị NA bằng giá trị trung bình sử dụng dplyr

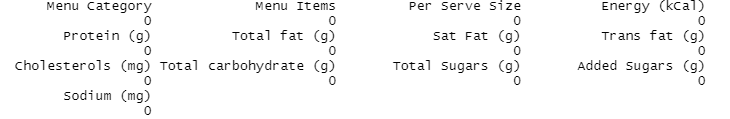
data <- data %>%

mutate(`Sodium (mg)` = ifelse(is.na(`Sodium (mg)`), mean(`Sodium (mg)`, na.rm = TRUE), `Sodium (mg)`))

# Kiểm tra lại số lượng giá trị NA trong từng cột

colSums(is.na(data))

Kết quả thực hiện là:



*Bảng 2.6: Kiểm tra số lượng NA của các biến*

## **2.3.THỐNG KÊ MÔ TẢ**

### **2.3.1. Tính toán các thống kê cơ bản**

Trước khi tiến hành các phân tích chi tiết và phức tạp hơn, việc tính toán các thống kê cơ bản là bước quan trọng nhằm hiểu rõ các đặc điểm chính của bộ dữ liệu.Những thống kê này cung cấp một cái nhìn tổng quan về phân phối và phạm vi của dữ liệu, giúp chúng ta xác định các giá trị trung bình, độ lệch chuẩn, phương sai và các phân vị quan trọng.

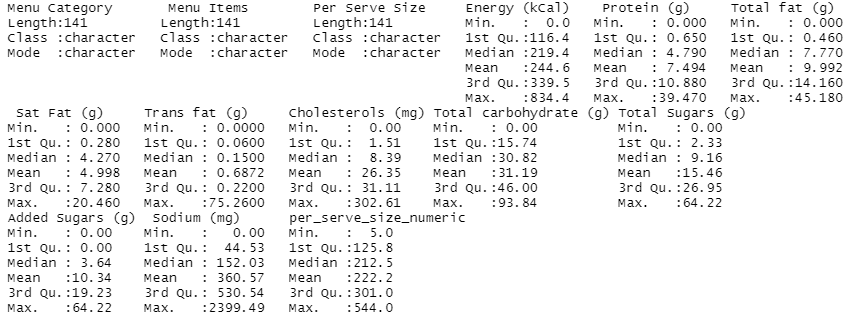
***Ví dụ***: Tính giá trị min, max, trung bình, phân vị 1/4, trung vị, phân vị 3/4, độ lệch chuẩn, phương sai về các biến sau:

* Energy (kCal)
* Protein (g)
* Total fat (g)
* Sat Fat (g)
* Trans fat (g)
* Cholesterols (mg)
* Total carbohydrate (g)
* Total Sugars (g)
* Added Sugars (g)
* Sodium (mg)

Ta sử dụng lệnh tổng quan ‘summary(data)’ để tóm tắt các giá trị trên:

data\_summary <- summary(data)

print(data\_summary)



*Bảng 2.7: Tóm tắt thống kê mô tả các nhóm*

***Ví dụ*** : Dựa vào bảng trên hãy đánh giá và nhận xét về biến “ Energy (kCal)”:

* Phân bố không đều: Sự khác biệt lớn giữa giá trị min (0 kCal) và max (834.4 kCal) cùng với việc trung bình cao hơn trung vị cho thấy rằng phân bố của biến năng lượng có khả năng lệch về phía các giá trị cao.
* Khoảng biến động lớn: Với khoảng biến động rộng từ 0 đến 834.4 kCal, menu này bao gồm cả các món ăn nhẹ nhàng và các món ăn rất giàu năng lượng.
* Sự đa dạng: Menu cung cấp một sự đa dạng lớn về lượng calo, điều này có thể đáp ứng nhu cầu của nhiều nhóm khách hàng khác nhau, từ những người cần kiểm soát calo chặt chẽ đến những người cần bổ sung năng lượng nhiều hơn.
* Món ăn giàu năng lượng: Một số món ăn có lượng calo rất cao, điều này có thể là lựa chọn tốt cho những người cần nhiều năng lượng như vận động viên, nhưng cũng cần chú ý vì có thể không phù hợp với những người cần hạn chế calo hoặc theo dõi sức khỏe.

Để có cái nhìn tổng quan về sự đa dạng và phong phú của các món ăn trong từng loại Menu, chúng ta cần vẽ biểu đồ số lượng món ăn theo từng danh mục Menu.

***Ví dụ***: Vẽ biểu đồ số lượng món ăn theo từng danh mục

Trước khi muốn vẽ biểu đồ, ta cần phải tải về các thư viện cần thiết như sau:

library(ggplot2)

library(dplyr)

Sau đó, chúng ta thực hiện việc tạo biểu đồ bằng hàm group\_by và summarise để nhóm dữ liệu theo Menu Category và tính số lượng món ăn trong mỗi nhóm.

# Tính số lượng của từng Menu Category

category\_counts <- data %>%

group\_by(`Menu Category`) %>%

summarise(count = n())

# Tạo biểu đồ

ggplot(category\_counts, aes(x = reorder(`Menu Category`, count), y = -count)) +

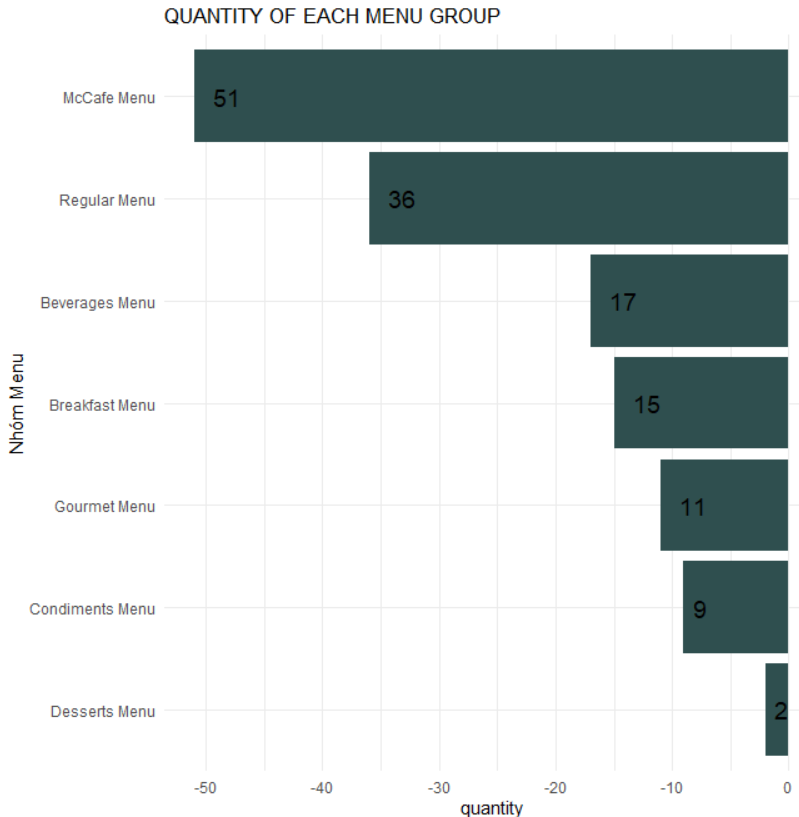
geom\_bar(stat = "identity", fill = "DarkSlateGray") +

geom\_text(aes(label = count), vjust = 0.5, hjust = -0.7, size = 5, color = "black") +

labs(title = "QUANTITY OF EACH MENU GROUP", x = "Nhóm Menu", y = "quantity") +

theme\_minimal() +

coord\_flip()



*Hình 2.1: Biểu đồ về số lượng món ăn theo Menu*

### **2.3.2. Công thức chung để tính toán các biến và vẽ biểu đồ**

**2.3.2.1. Tính trung bình các biến và vẽ biểu đồ thống kê**

Để hiểu rõ hơn về xu hướng và sự phân bố của các giá trị dinh dưỡng trong từng loại Menu, chúng ta sẽ tính toán giá trị trung bình của các biến. Sau đó, các giá trị trung bình này sẽ được trực quan hóa bằng biểu đồ thống kê, giúp chúng ta dễ dàng nhận thấy sự khác biệt và tương quan giữa các loại Menu.

Đoạn code dưới đây có thể được sử dụng để tính toán trung bình và vẽ biểu đồ cho bất kỳ biến nào trong dữ liệu; chỉ cần thay thế tên biến trong các hàm group\_by, summarize, và aes của ggplot2.

#Tính trung bình các biến

average\_energy <- data %>%

group\_by(`Menu Category`) %>%

summarize(mean\_VARIABLE\_NAME = mean(`VARIABLE\_NAME)`, na.rm = TRUE)) %>%

arrange(desc(mean\_VARIABLE\_NAME))

# Vẽ biểu đồ cột

ggplot(data, aes(x = factor(`Menu Category`, levels = average\_VARIABLE\_NAME$`Menu Category`), y = `TÊN BIẾN`, fill = `Menu Category`)) +

geom\_bar(stat = "summary", fun = "mean") +

geom\_text(stat = "summary", aes(label = round(..y.., 2)), vjust = -0.5) +

position = position\_dodge(width = 0.9)) +

labs(title = "Biểu đồ Thống kê trung bình VARIABLE\_NAME theo danh mục menu",

x = "Danh mục menu",

y = "VARIABLE\_NAME") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

Trong đoạn code trên:

* Thay VARIABLE\_NAME bằng tên biến mà bạn muốn tính trung bình và vẽ biểu đồ, ví dụ như Energy (kCal), Protein (g), Total fat (g). v.v.

***Ví dụ***: Hãy tính toán trung bình và vẽ biểu đồ biến Total fat (g) sau đó đưa ra nhận xét.

Đoạn code dưới đây được thay VARIABLE\_NAME bằng tên biến Total fat (g)

# Tính trung bình total fat theo danh mục menu và sắp xếp giảm dần

average\_total\_fat <- data %>%

group\_by(`Menu Category`) %>%

summarize(mean\_total\_fat = mean(`Total fat (g)`, na.rm = TRUE)) %>%

arrange(desc(mean\_total\_fat))

# Vẽ biểu đồ cột

ggplot(data, aes(x = factor(`Menu Category`, levels = average\_total\_fat$`Menu Category`), y = `Total fat (g)`, fill = `Menu Category`)) +

geom\_bar(stat = "summary", fun = "mean") +

geom\_text(stat = "summary", aes(label = round(..y.., 2)), vjust = -0.5,

position = position\_dodge(width = 0.9)) +

labs(title = "Biểu đồ Thống kê trung bình lượng Total fat theo danh mục menu",

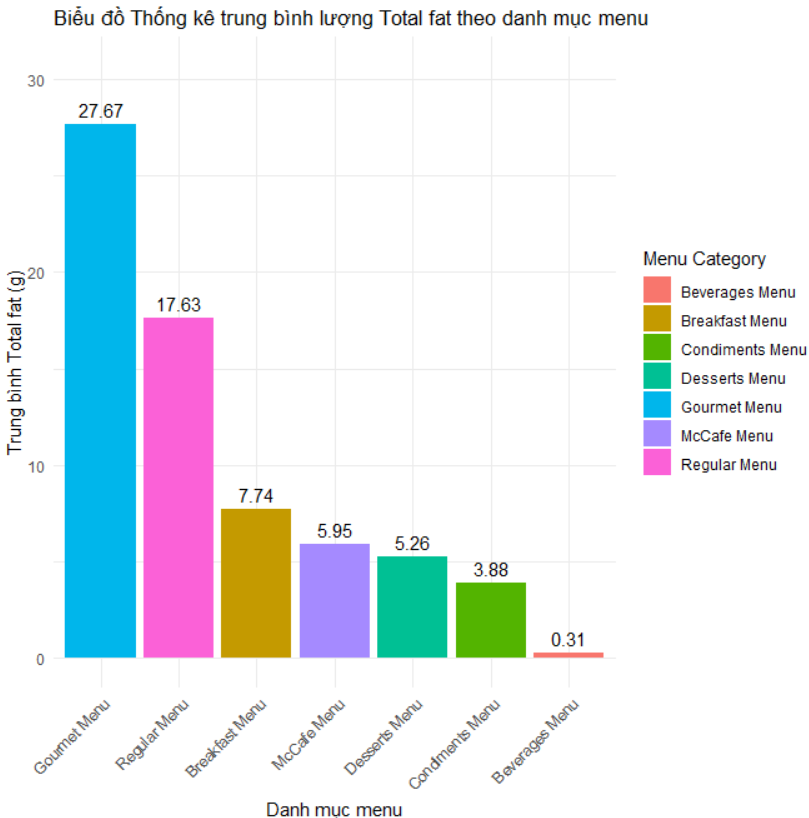
x = "Danh mục menu",

y = "Trung bình Total fat (g)") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

Biểu đồ sau khi được chạy ra, nhận được kết quả là:



*Hình 2.2: Biểu đồ thống kê trung bình lượng Total fat theo Menu*

Biểu đồ này thể hiện trung bình lượng chất béo tổng hợp theo từng danh mục menu. Dưới đây là nhận xét về biểu đồ:

* Nhóm Menu Gourmet: Có lượng chất béo trung bình cao nhất, đạt 27.67g. Đây có thể là các món ăn cao cấp hoặc đặc biệt.
* Nhóm Menu Regular: Có lượng chất béo trung bình là 17.63g. Đây có thể là các món ăn phổ biến và tiêu chuẩn.
* Nhóm Menu Breakfast và nhóm Menu McCafe: Có giá trị tương tự, khoảng 6-7g (lần lượt là 7.74 và 5.95). Có thể là cái loại đồ ăn sáng hoặc đồ uống từ McCafe.
* Nhóm Menu Desserts,nhóm Menu Condiments và nhóm Menu Beverages : Có lượng chất béo trung bình thấp nhất, lần lượt là 5.26, 3.88 và 0.31g. Đây có thể là các món tráng miệng, các loại gia vị, phụ gia hoặc đồ uống không chứa nhiều chất béo.

**2.3.2.2. Biểu đồ phân phối theo các nhóm trong Menu Category.**

Để có cái nhìn sâu sắc hơn về sự phân bố của các giá trị dinh dưỡng trong từng nhóm của Menu Category, chúng ta sẽ vẽ biểu đồ phân phối cho các nhóm này. Biểu đồ phân phối sẽ giúp chúng ta nhận thấy rõ hơn về sự khác biệt và đặc trưng của mỗi nhóm, từ đó cung cấp thông tin quan trọng cho việc phân tích và so sánh các loại món ăn trong thực đơn.

Đoạn mã này cung cấp một công cụ linh hoạt để khám phá phân phối của các biến trong tập dữ liệu. Bạn có thể sử dụng mã này để tạo ra biểu đồ phân phối cho bất kỳ biến nào, từ năng lượng tiêu thụ đến lượng chất béo, chỉ cần thay đổi tên biến trong các hàm aes và labs của ggplot2. Điều này giúp bạn nhanh chóng và dễ dàng thí nghiệm với nhiều biến và hiểu rõ hơn về phân phối của chúng trong các nhóm danh mục khác nhau.

Dưới đây là đoạn code minh họa:

ggplot(data, aes(x = `VARIABLE\_NAME`)) +

geom\_density(aes(fill = `Menu Category`), alpha = 0.5) +

facet\_wrap(~ `Menu Category`) +

labs(title = "Phân phối VARIABLE\_NAME theo nhóm Menu Category", x =

"VARIABLE\_NAME", y = "Mat do") +

theme\_minimal()

Để hiểu rõ hơn về đoạn code này, ta có ví dụ sau:

***Ví dụ*:** Hãy vẽ biểu đồ phân phối của Total carbohydrate và đưa ra nhận xét.

Áp dụng theo đoạn code chung ở trên, ta thay VARIABLE\_NAME bằng tên biến Total carbohydrate (g):

ggplot(data, aes(x = `Total carbohydrate (g)`)) +

geom\_density(aes(fill = `Menu Category`), alpha = 0.5) +

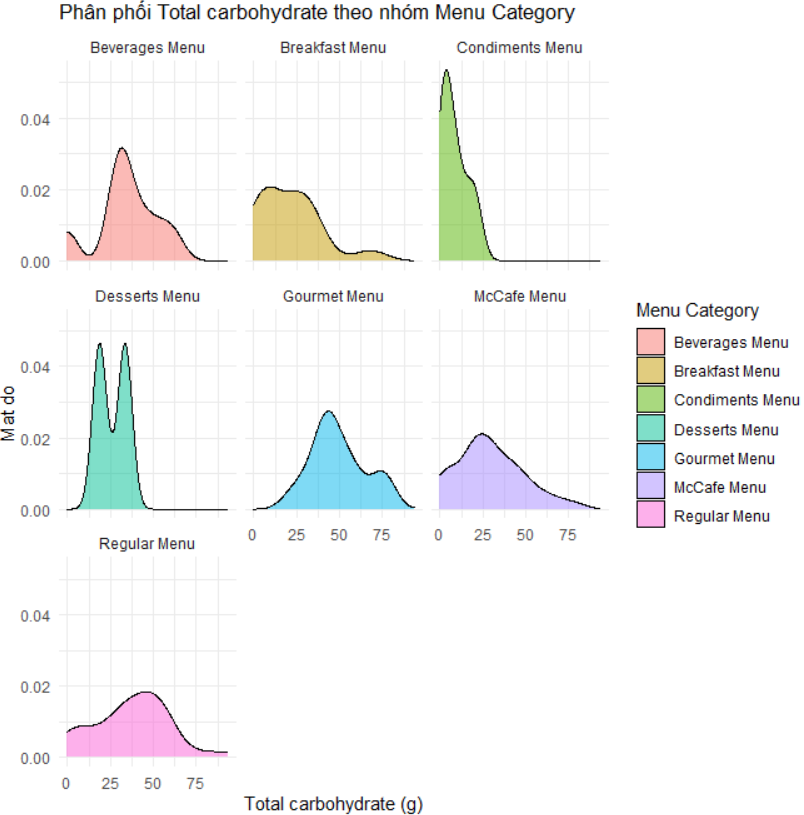
facet\_wrap(~ `Menu Category`) +

labs(title = "Phân phối Total carbohydrate theo nhóm Menu Category", x =

"Total carbohydrate (g)", y = "Mat do") +

theme\_minimal()

Biểu đồ nhận được hình ảnh là:



*Hình 2.3: Biểu đồ phân phối của Total carbohydrate*

* **Nhận xét:**
* Nhóm Beverages Menu:
* Phân phối rộng với nhiều mức carbohydrate khác nhau, từ rất thấp đến khoảng 60g.
* Mật độ cao nhất nằm ở khoảng 25-35g sau đó giảm dần.
* Nhóm Breakfast Menu:
* Có hai đỉnh phân phối chính, một ở khoảng 0-10g, một ở khoảng 25-35g.
* Mức carbohydrate thay đổi khá đa dạng.
* Nhóm Condiments Menu:
* Phân phối chủ yếu tập trung ở mức rất thấp, dưới 20g.
* Hầu như không có món nào vượt quá 30g carbohydrate.
* Nhóm Desserts Menu:
* Có hai đỉnh rất rõ ràng, một ở khoảng 0-15g và một ở khoảng 40-50g.
* Điều này cho thấy sự khác biệt lớn giữa các loại tráng miệng, một số rất ít carbohydrate và một số rất cao.
* Nhóm Gourmet Menu:
* Phân phối rộng với một đỉnh rõ ràng ở khoảng 25-35 g.
* Có sự biến động lớn về lượng carbohydrate trong nhóm này.
* Nhóm McCafe Menu:
* Phân phối tương đối thấp và hẹp, tập trung chủ yếu dưới 25g.
* Mật độ cao nhất là ở khoảng 15-25g.
* Nhóm Regular Menu:
* Phân phối rộng với nhiều mức carbohydrate khác nhau, từ rất thấp đến khoảng 75g.
* Có một đinh đỉnh rõ ràng ở khoảng 10-15g và một đỉnh nhỏ hơn ở khoảng 30-45g.

**2.3.2.3. Vẽ biểu đồ phân phối ECDF**

Biểu đồ phân phối Empirical Cumulative Distribution Function (ECDF) là một công cụ mạnh mẽ để khám phá và mô tả phân phối của dữ liệu. ECDF hiển thị phần trăm quan sát có giá trị nhỏ hơn hoặc bằng một giá trị cụ thể trên trục x. Bằng cách sử dụng ECDF, chúng ta có thể nhìn thấy mức độ phân phối của dữ liệu trên toàn bộ phạm vi, từ giá trị nhỏ nhất đến lớn nhất. Điều này giúp chúng ta đánh giá mức độ phân tán và phân phối của dữ liệu một cách trực quan và dễ hiểu.

Dưới đây là ví dụ minh họa về biến Energy (kCal), ngoài ra ta có thể thay bằng các biến khác.

***Ví dụ***: Vẽ biểu đồ phân phối ECDF, từ đó đưa ra nhận xét và đánh giá.

# Tạo biến thể hiện phân phối ECDF cho (`Energy (kCal)`) theo từng danh mục menu

ecdf\_data <- data %>%

group\_by(`Menu Category`) %>%

mutate(Energy\_ECDF = ecdf(`Energy (kCal)`)(`Energy (kCal)`))

# Vẽ biểu đồ phân phối ECDF

data %>%

group\_by(`Menu Category`) %>%

mutate(Energy\_ECDF = ecdf(`Energy (kCal)`)(`Energy (kCal)`)) %>%

ggplot(aes(x = `Energy (kCal)`, y = Energy\_ECDF, color = `Menu Category`)) +

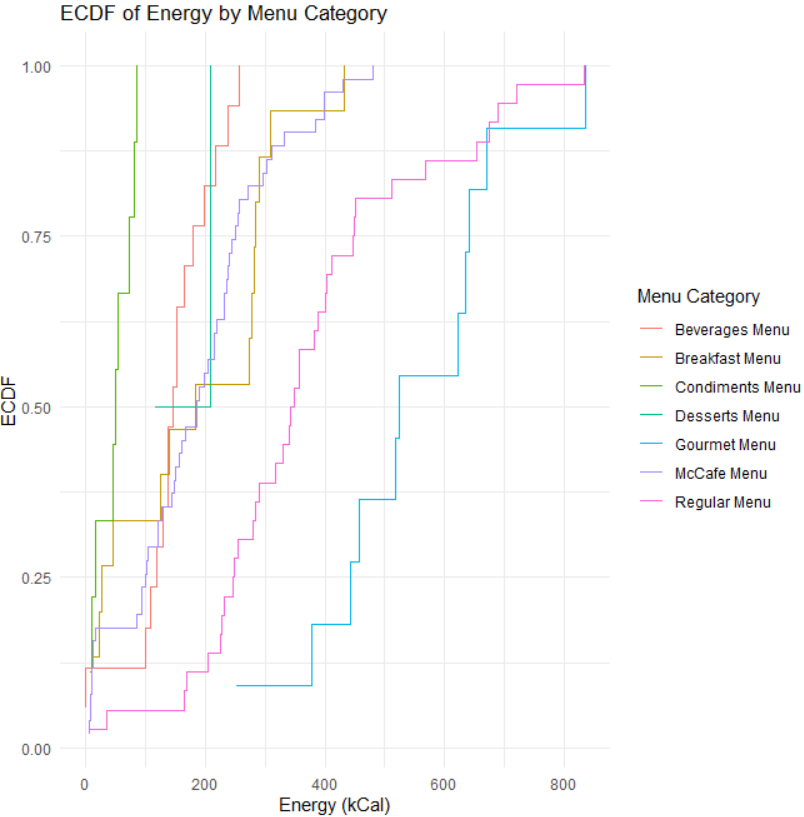
geom\_step() +

labs(title = "ECDF of Energy by Menu Category",

x = "Energy (kCal)",

y = "ECDF") +

theme\_minimal()



*Hình 2.4: Biểu đồ phân phối ECDF của biến Energy (kCal)*

* **Nhận xét:**
* Nhóm Beverages:
* Đường ECDF khá dốc và kết thúc ở mức năng lượng khoảng 250kCal.
* Phần lớn các món trong nhóm này có mức năng lượng dưới 200kCal.
* Nhóm Breakfast Menu:
* Đường ECDF khá dốc và đạt giá trị ở mức khoảng 400kCal.
* Khoảng 75% các món ăn sáng có năng lượng dưới 300kCal.
* Nhóm Condiments Menu:
* Đường ECDF rất dốc và đạt giá trị ở mức năng lượng khoảng 100kCal.
* Hầu hết các gia vị có năng lượng rất thấp.
* Nhóm Desserts Menu:
* Đường ECDF dốc và đạt giá trị ở mức năng lượng khoảng 200kCal.
* Nhóm Gourmet Menu:
* Đường ECDF tăng đều đặn và đạt giá trị ở mức khoảng 800 kCal.
* Phần lớn các món cao cấp có năng lượng dưới 650kCal.
* Nhóm McCafe Menu:
* Đường ECDF khá dốc và đạt giá trị ở mức khoảng 450kCal.
* Khoảng 75% các món trong McCafe có năng lượng dưới khoảng 400kCal.
* Nhóm Regular Menu:
* Đường ECDF trải dài từ 0 đến 800 kCal.
* Khoảng 50% các món món thường có năng lượng dưới 400kCal và 75% dưới 600kCal.
* **Đánh giá:**
* Phân phối năng lượng: Nhóm Condiments Menu có năng lượng thấp nhất, phù hợp với tính chất của gia vị thường không đóng góp nhiều năng lượng. Nhóm Beverages Menu và McCafe Menu cũng có năng lượng thấp đến trung bình, trong khi các nhóm Breakfast Menu, Gourmet Menu và Regular Menu thường có sự phân bố rộng hơn.
* Nhóm Regular Menu có sự phân phối rộng nhất, với một số món có mức năng lượng rất cao lên đến 800kCal, phản ánh đặc điểm của các món giàu năng lượng.
* Nhìn chung, các nhóm có sự khác biệt rõ rệt về mức năng lượng. Nhóm Condiments Menu có mức năng lượng thấp nhất, trong khi nhóm Gourmet Menu và Regular Menu có mức năng lượng cao nhất.

## **2.4. PHÂN TÍCH MỐI QUAN HỆ GIỮA CÁC BIẾN**

Để hiểu rõ hơn về cách các biến dinh dưỡng trong thực đơn của McDonald's tương tác và ảnh hưởng lẫn nhau, chúng ta sẽ tiến hành phân tích mối quan hệ giữa các biến.

Bằng cách sử dụng các công cụ thống kê và trực quan hóa dữ liệu, chúng ta có thể xác định các mối tương quan, xu hướng và tương tác giữa các yếu tố dinh dưỡng như năng lượng, chất béo, protein, carbohydrate và các thành phần khác.

Phân tích này sẽ cung cấp cái nhìn sâu sắc về cấu trúc dinh dưỡng của thực đơn, giúp đưa ra những quyết định dựa trên dữ liệu để cải thiện chất lượng dinh dưỡng và sự cân bằng của các món ăn.

### **2.4.1. Tính ma trận tương quan**

Để khám phá mối quan hệ giữa các biến dữ liệu, chúng ta sẽ bắt đầu bằng việc tính ma trận tương quan. Ma trận tương quan cung cấp cái nhìn toàn diện về mức độ và hướng của mối quan hệ tuyến tính giữa các biến.

Để tính ma trận tương quan trong R, ta dùng cú pháp sau:

**cor\_matrix <- cor(data)**

***Ví dụ:*** Hãy tính ma trận tương quan và vẽ biểu đồ về các biến dinh dưỡng.

Trước hết, ta cần tải các thư viện sau để thuận tiện hơn trong việc vẽ biểu đồ.

library(ggplot2)

library(reshape2)

Sau đó, ta tiến hành tính toán ma trận tương quan và vẽ biểu đồ.

# Lọc ra các cột dữ liệu số

numeric\_data <- data[,sapply(data, is.numeric)]

# Tính ma trận tương quan

correlation\_matrix <- cor(numeric\_data)

# Chuyển đổi ma trận tương quan sang dạng dataframe

correlation\_df <- as.data.frame(as.table(correlation\_matrix))

names(correlation\_df) <- c("Variable1", "Variable2", "Correlation")

# Vẽ biểu đồ heatmap

ggplot(correlation\_df, aes(x = Variable1, y = Variable2, fill = Correlation)) +

geom\_tile() +

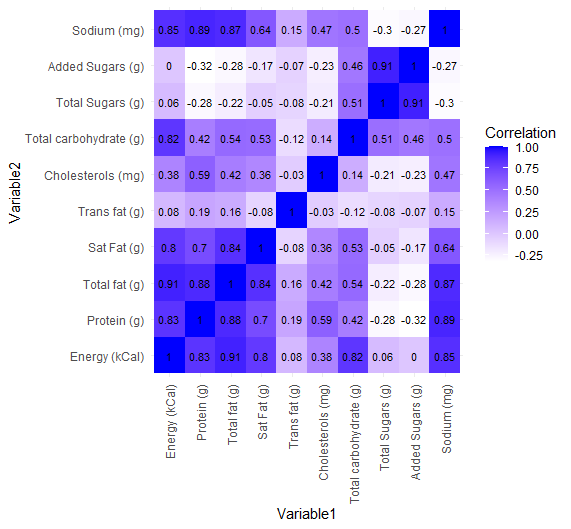
geom\_text(aes(label = round(Correlation, 2)), color = "black", size = 3) +

scale\_fill\_gradient(low = "white", high = "blue") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust=1))

Kết quả là:



*Hình 2.5: Ma trận tương quan giữa các thành phần dinh dưỡng*

**Nhận xét :**

* Mối tương quan tích cực: Năng lượng (KCal), Protein (g),
* Total Fat (g) và nhiều yếu tố khác có mối tương quan tích cực với nhau (ô màu đậm).Điều này có nghĩa là khi một yếu tố tăng, các yếu tố khác cũng có xu hướng tăng theo( tỉ lệ thuận).
* Mối tương quan tiêu cực: Có mối tương quan tiêu cực giữa Added Sugars (lượng đường thêm vào) và các yếu tố khác (ô màu xám).
* Điều này có thể chỉ ra rằng việc tiêu thụ đường thêm vào có thể ảnh hưởng đến các yếu tố khác trong chế độ ăn uống., tăng các chỉ số tiêu cực.

### **2.4.2. Trực quan hóa mối liên hệ giữa các biến**

Trực quan hóa mối liên hệ giữa các biến là một phần quan trọng của quá trình phân tích dữ liệu. Điều này giúp chúng ta hiểu rõ hơn về cách các biến tương tác và ảnh hưởng lẫn nhau trong tập dữ liệu.

**2.4.2.1. Ma trận tương quan giữa các biến và cách mô tả**

*2.4.2.1.1. Khái niệm*

Ma trận tương quan là một công cụ quan trọng trong phân tích dữ liệu để xác định mối quan hệ tương quan giữa các biến. Đây là một bảng số liệu với các giá trị tương quan được tính giữa tất cả các cặp biến trong tập dữ liệu. Mỗi phần tử trong ma trận tương quan thể hiện mức độ tương quan giữa hai biến, thường từ -1 đến 1.

*2.4.2.1.2. Đặc điểm:*

Mô tả của ma trận tương quan thường gồm những đặc điểm sau:

* Phạm vi giá trị: Giá trị của tương quan nằm trong khoảng từ -1 đến 1. Giá trị gần -1 hoặc 1 cho thấy một mối quan hệ tương quan mạnh mẽ giữa các biến, trong khi giá trị gần 0 cho thấy không có hoặc có mối quan hệ yếu.
* Hướng của mối quan hệ: Dấu của giá trị tương quan (dương hoặc âm) cho biết hướng của mối quan hệ. Một giá trị dương cho thấy mối quan hệ tương quan tích cực, nghĩa là khi một biến tăng, biến kia cũng tăng (và ngược lại). Một giá trị âm cho thấy mối quan hệ tương quan tiêu cực, nghĩa là khi một biến tăng, biến kia giảm (và ngược lại).
* Sức mạnh của mối quan hệ: Giá trị tương quan càng gần -1 hoặc 1, mối quan hệ càng mạnh. Các giá trị gần 0 cho thấy mối quan hệ yếu.
* Biểu đồ ma trận tương quan: Một biểu đồ ma trận tương quan có thể được sử dụng để trực quan hóa ma trận tương quan. Trên biểu đồ này, mỗi ô biểu thị một cặp biến và màu sắc hoặc kích thước của ô thể hiện mức độ tương quan.

*2.4.2.1.3. Tính toán ma trận và vẽ biểu đồ*

1. *Biểu đồ tròn*

Đầu tiên, chúng ta sẽ tính toán ma trận tương quan từ dữ liệu. Sau đó, chúng ta vẽ biểu đồ tròn để trực quan hóa các giá trị tương quan này. Biểu đồ tròn sẽ cho chúng ta cái nhìn tổng quát về sức mạnh và hướng của mối quan hệ giữa các biến.

library(readr)

library(dplyr)

library(corrplot)

for (col in names(data)) {

col\_median <- median(data[[col]], na.rm = TRUE) # Tính trung vị của cột

data[[col]][is.na(data[[col]])] <- col\_median # Gán trung vị vào các giá trị trống

}

# Hiển thị data frame sau khi thay thế

print(col\_median)

print("Data frame sau khi thay thế:")

print(data)

# Lọc ra chỉ các cột có dữ liệu số

numeric\_columns <- data %>%

select\_if(is.numeric)

# Tính ma trận tương quan

ma\_tran <- cor(numeric\_columns)

# Biểu đồ tròn tương quan

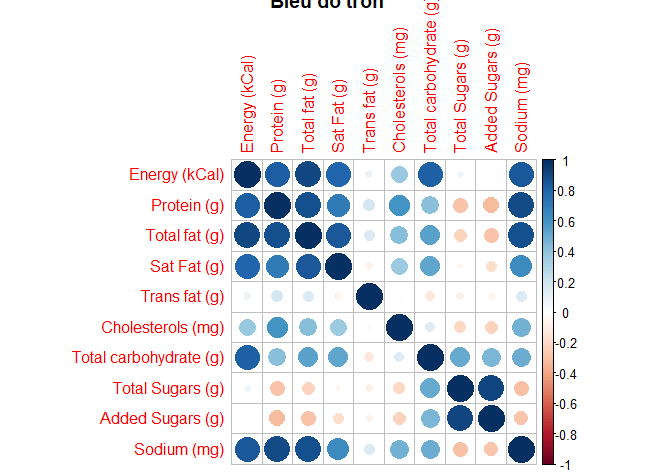
corrplot(ma\_tran,

method = "circle",

type = "full",

title = "Biểu đồ tròn")

Kết quả thực hiện là:



*Hình 2.6: Ma trận tương quan thông qua biểu đồ tròn và màu sắc*

**Nhận xét:**

* Mối tương quan giữa "Total fat (g)" và "Energy (kCal)" rất cao (gần như là 1), cho thấy rằng khi lượng chất béo tổng tăng thì năng lượng cũng tăng."
* Total fat (g)", "Sat Fat (g)", "Trans fat (g)", và "Cholesterol (mg)" đều có mối tương quan mạnh với Energy (kcal), cho thấy rằng các chất béo và cholesterol đóng góp đáng kể vào lượng năng lượng của thực phẩm.

1. *Biểu đồ số*

Biểu đồ số sẽ hiển thị phân phối của mỗi biến, giúp chúng ta nhận biết các phân vị của phân phối.

# ma trận tương quan số

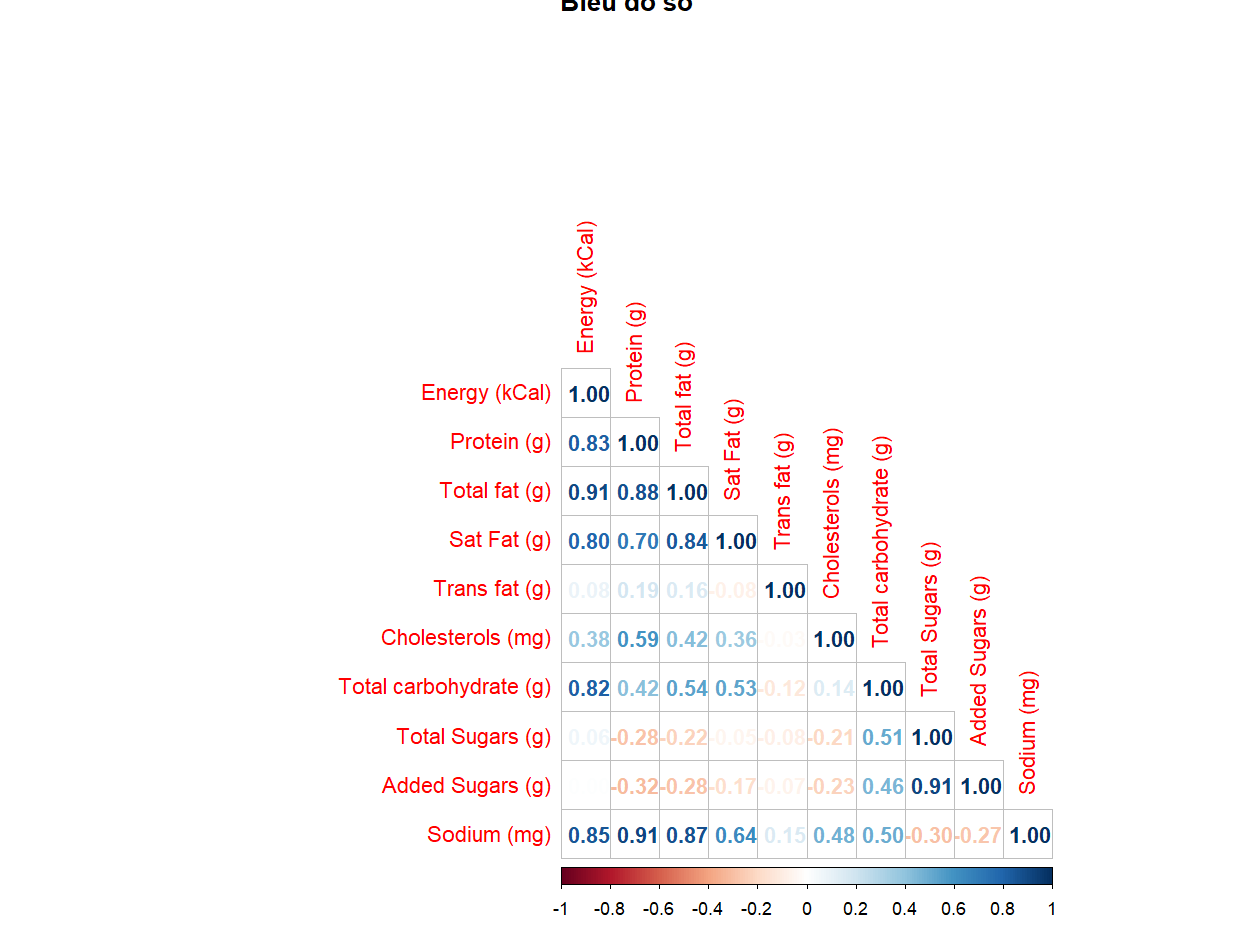
corrplot(ma\_tran,

method = “number”,

type = “lower”,

title = “Biểu đồ số”)

Kết quả là:



*Hình 2.7: Ma trận tương quan thông qua số liệu và màu sắc*

**Nhận xét:**

* Total fat (g) tương quan mạnh với "Sat Fat (g)" (0.70) và "Trans fat (g)" (0.59).
* Sodium (mg) có mối tương quan khá cao với "Total fat (g)" (0.91) và "Protein (g)" (0.87).
* "Total Sugars (g)" và "Added Sugars (g)" có mối tương quan gần như tuyệt đối (0.91), cho thấy rằng phần lớn lượng đường tổng trong thực phẩm đến từ đường thêm vào.
* "Total fat (g)" và "Sat Fat (g)" cũng có mối tương quan cao (0.70), phản ánh rằng thực phẩm giàu chất béo tổng thường cũng có nhiều chất béo bão hòa.

*c) Biểu đồ màu*

Màu sắc sẽ được sử dụng để thể hiện mức độ tương quan, với các giá trị tương quan cao được biểu diễn bằng màu sắc tương phản cao hơn.

# ma trận tương quan màu sắc

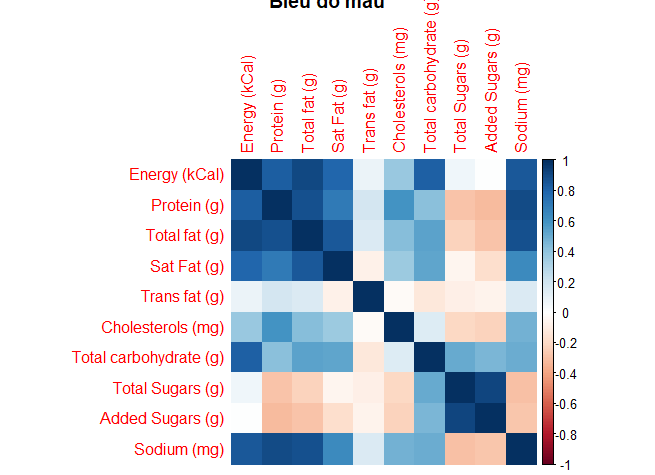
corrplot(ma\_tran,

method ="color",

type = "full",

title = "Biểu đồ màu")

Kết quả là:



*Hình 2.8: Ma trận tương quan thể hiện bởi màu sắc*

**Nhận xét:**

* Total Sugars (g) và Added Sugars (g): Mối tương quan cực kỳ cao, gần như tuyệt đối.
* Cholesterols (mg) có mối tương quan thấp với nhiều yếu tố khác, đặc biệt là Total Sugars (g) và Added Sugars (g).
* Total carbohydrate (g) có mối tương quan không quá mạnh với hầu hết các yếu tố khác, ngoại trừ một số yếu tố như Total Sugars (g) và Added Sugars (g), nơi có một chút tương quan dương.

*d) Biểu đồ bánh*

Biểu đồ bánh dùng để trực quan hóa phần trăm phân phối của mỗi biến trong tập dữ liệu, giúp chúng ta hiểu rõ hơn về sự đa dạng của dữ liệu và cách mỗi biến đóng góp vào tổng thể.

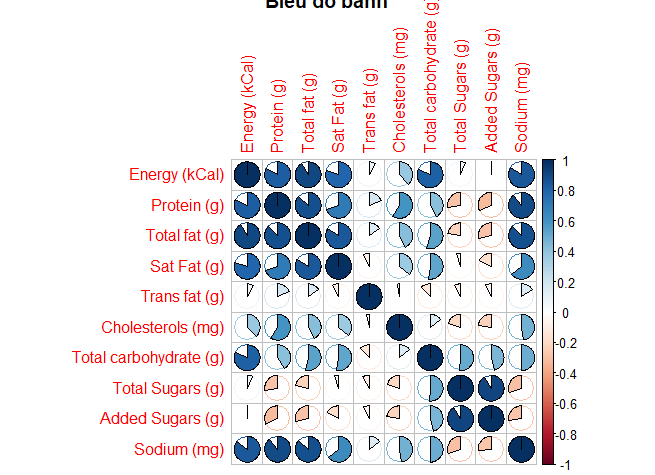
# ma trận tương quan bánh tròn

corrplot(ma\_tran,

method ="pie",

type = "full",

title = "Biểu đồ bánh")



*Hình 2.9: Ma trận tương quan biểu diễn thông qua biểu đồ bánh*

**Nhận xét:**

* Total Sugars (g) và Added Sugars (g) có mối tương quan cực kỳ cao.
* Cholesterols (mg) có mối tương quan yếu với nhiều yếu tố khác, đặc biệt là Total Sugars (g) và Added Sugars (g).
* Total carbohydrate (g) có mối tương quan dương nhẹ với Total Sugars (g) và Added Sugars.

**2.4.2.2. Biểu đồ Scatter thể hiện các cặp số tương quan**

Biểu đồ scatter plot (hay còn gọi là biểu đồ phân tán) là một loại biểu đồ dùng để hiển thị mối quan hệ giữa hai biến số. Mỗi điểm trên biểu đồ đại diện cho một quan sát trong tập dữ liệu với giá trị của một biến được biểu diễn trên trục x và giá trị của biến kia được biểu diễn trên trục y. Biểu đồ scatter plot rất hữu ích trong việc nhận diện mối tương quan giữa các biến và phát hiện các mẫu, xu hướng hoặc giá trị ngoại lệ trong dữ liệu.

1. *Tính các cặp có giá trị tương quan cao nhất*

Từ tập dữ liệu “India\_Menu.csv”, ta chọn ra các cặp giá trị dinh dưỡng có mức độ tương quan cao nhất. Các cặp giá trị này có mối quan hệ chặt chẽ với nhau và quy định giá trị dinh dưỡng có trong từng sản phẩm.

# Lấy chỉ số của top 10 cặp giá trị tương quan cao nhất

top\_indices <- which(upper.tri(cor\_matrix, diag = FALSE), arr.ind = TRUE)[order(cor\_matrix[upper.tri(cor\_matrix, diag = FALSE)], decreasing = TRUE)[1:10], ]

# In ra top 10 cặp biến và giá trị tương quan tương ứng

for (i in 1:10) {

row\_index <- top\_indices[i, 1]

col\_index <- top\_indices[i, 2]

correlation <- cor\_matrix[row\_index, col\_index]

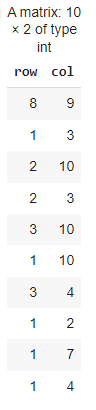
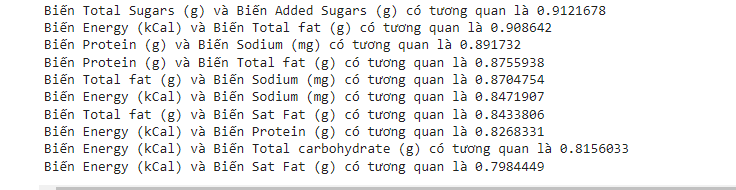
cat("Biến", rownames(cor\_matrix)[row\_index], "và",

"Biến", colnames(cor\_matrix)[col\_index], "có tương quan là",

correlation, "\n")

}

Kết quả nhận được là:



*Bảng 2.8: Các biến có giá trị tương quan cao nhất*

***Ví dụ:*** Dựa vào bảng 2.8, hãy vẽ biểu đồ Scatter plot giữa hai biến Energy (kCal) và biến Total fat (g) để thể hiện mối quan hệ tương quan, và đưa ra nhận xét.

# Vẽ biểu đồ scatter plot

ggplot(data, aes(x = `Energy (kCal)`, y = `Total fat (g)`, color = `Menu Category`)) +

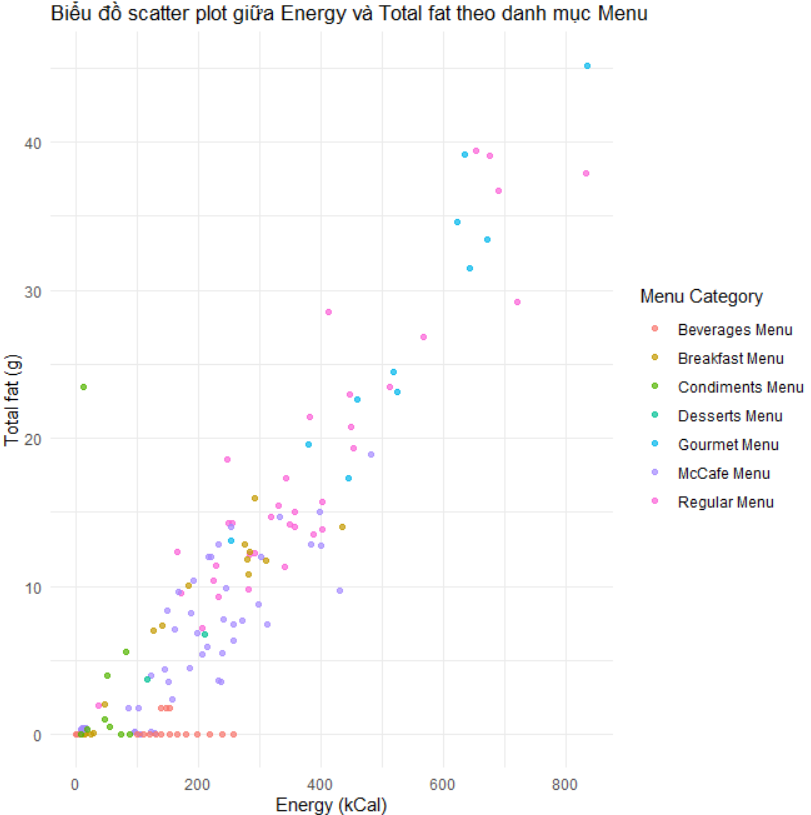
geom\_point(alpha = 0.7) +

labs(title = "Biểu đồ scatter plot giữa Energy và Total fat theo danh mục Menu",

x = "Energy (kCal)",

y = "Total fat (g)") +

theme\_minimal()



*Hình 2.10: Biểu đồ Scatter plot giữa Energy (kCal) và Total fat (g)*

* **Nhận xét:**
* Phân tán dữ liệu:
* Dữ liệu trải dài từ 0 đến khoảng 800Kcal cho năng lượng và từ 0 đến khoảng 45g cho tổng lượng chất béo.
* Có một số điểm dữ liệu nằm riêng biệt ở phía trên của biểu đồ, cho thấy các món ăn có hàm lượng năng lượng và chất béo cao.
* So sánh giữa các danh mục:
* Beverages Menu: Hầu hết các điểm tập trung ở mức năng lượng và chất béo thấp.
* Breakfast Menu và Regular Menu: Có sự phân bố rộng hơn, với một số món ăn có hàm lượng năng lượng và chất béo cao.
* Condiments Menu và McCafe Menu: Có một số điểm với mức năng lượng và chất béo cao, nhưng đa số nằm ở mức trung bình.
* Gourmet Menu: Phân bố không đồng đều, nhưng không có nhiều món ăn có hàm lượng năng lượng và chất béo cực cao.
* Điểm nổi bật:
* Một số điểm dữ liệu của Regular Menu và Desserts Menu có hàm lượng năng lượng và chất béo rất cao, cho thấy sự đa dạng lớn trong các món ăn thuộc những danh mục này.
* Condiments Menu có hàm lượng chất béo và năng lượng rất thấp, có thể do đặc thù của các món gia vị.
* Kết luận:
* Biểu đồ này giúp nhận biết rõ mối quan hệ giữa năng lượng và tổng lượng chất béo trong các món ăn thuộc các danh mục menu khác nhau.
* Thực khách có thể sử dụng thông tin này để lựa chọn các món ăn phù hợp với nhu cầu dinh dưỡng của mình, đặc biệt là khi quan tâm đến lượng chất béo và năng lượng tiêu thụ.

1. *Tính các cặp có giá trị tương quan thấp nhất*

Cũng tương tự như phần *a)*, phần *b)* này ta cũng tính các cặp giá trị tương quan thấp nhất và xem chúng có mối quan hệ với nhau như thế nào.

# Chọn ra các biến có hệ số tương quan nhỏ nhất và trực quan chúng

bottom\_indices <- which(upper.tri(correlation\_matrix, diag = FALSE), arr.ind = TRUE)[order(correlation\_matrix[upper.tri(correlation\_matrix, diag = FALSE)], decreasing = FALSE)[1:5], ]

# Lấy chỉ số của top 5 cặp giá trị tương quan nhỏ nhất

for (i in 1:5) {

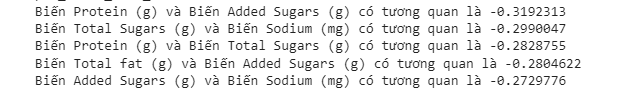
row\_index <- bottom\_indices[i, 1]

col\_index <- bottom\_indices[i, 2]

correlation <- correlation\_matrix[row\_index, col\_index]

cat("Biến", rownames(correlation\_matrix)[row\_index], "và",

"Biến", colnames(correlation\_matrix)[col\_index], "có tương quan là", correlation, "\n")}



*Bảng 2.9: Các biến có giá trị tương quan thấp nhất*

***Ví dụ***: Nhìn vào Bảng 2.9, hãy vẽ biểu đồ tán xạ giữa 2 biến Added Sugars (g) và biến Sodium (mg), và rút ra nhận xét.

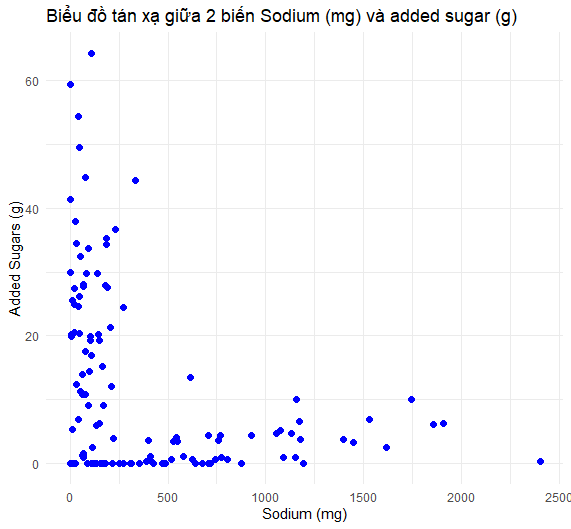
Ta có: Biến sodium và Biến added sugar (g) có tương quan là 0.5384779

ggplot(data, mapping = aes(x = `Sodium (mg)`, y = `Added Sugars (g)`))+

geom\_point(color = "blue", size = 2, alpha = 1)+

theme\_minimal() +

labs(title = "Biểu đồ tán xạ giữa 2 biến Sodium (mg) và added sugar (g) ")



*Hình 2.11: Biểu đồ thể hiện mối liên hệ giữa Sodium và Added Sugars*

**Nhận xét:**

* Có một cụm dày đặc các điểm dữ liệu tập trung ở phần dưới của thang đo sodium, cho thấy mối quan hệ giữa natri và lượng đường được thêm vào trong một số mẫu hoặc sản phẩm đang được phân tích.
* Khi di chuyển qua phải theo trục ngang, cho thấy sự tăng lên về lượng natri, có một xu hướng nhẹ tăng về lượng đường được thêm vào. Tuy nhiên, đây không phải là mối quan hệ mạnh hoặc rõ ràng.

# **CHƯƠNG 3 : TRỰC QUAN HÓA DỮ LIỆU**

## **3.1. GIỚI THIỆU VỀ TRỰC QUAN HÓA DỮ LIỆU**

### **3.1.1.Khái niệm**

Trực quan hóa dữ liệu là quá trình chuyển đổi dữ liệu từ các định dạng số liệu hoặc văn bản sang các định dạng đồ họa, như biểu đồ, đồ thị, bản đồ hoặc các biểu đồ trực quan khác. Mục tiêu của trực quan hóa dữ liệu là giúp người xem dễ dàng hiểu và phân tích thông tin từ dữ liệu, nhận ra các xu hướng, mô hình, và mối quan hệ trong dữ liệu một cách nhanh chóng và trực quan.

### **3.1.2. Mục đích**

* Hiểu rõ dữ liệu: Trực quan hóa giúp phát hiện các mẫu, xu hướng và mối quan hệ mà có thể không rõ ràng từ dữ liệu thô.
* Giao tiếp thông tin: Giúp truyền đạt thông tin và kết quả phân tích một cách hiệu quả đến người dùng cuối, đồng nghiệp, hoặc các bên liên quan.

### **3.1.3. Phân loại**

* Biểu đồ Bar chart: Thể hiện số lượng hoặc giá trị của các mục khác nhau.
* Biểu đồ Line chart: Thể hiện xu hướng của một hoặc nhiều biến qua thời gian.
* Biểu đồ Pie chart: Thể hiện tỷ lệ phần trăm của các phần trong tổng thể.
* Biểu đồ Scatter plot: Thể hiện mối quan hệ giữa hai biến số.
* Biểu đồ Boxplot: Thể hiện phân bố của một biến và các giá trị ngoại lai.
* Bản đồ Heatmap: Thể hiện cường độ dữ liệu qua màu sắc.

## **3.2. BIỂU ĐỒ VỀ TRỰC QUAN HÓA DỮ LIỆU**

### **3.2.1. Biểu đồ tròn phân phối mục theo danh sách thực đơn**

library(ggplot2)

# Dữ liệu

data1 <- data.frame(

Menu\_Category = c("Regular Menu", "Breakfast Menu", "McCafe Menu", "Desserts Menu", "Gourmet Menu", "Beverages Menu", "Condiments Menu"),

Frequency = c(63, 15, 38, 3, 18, 36, 4) # Số lượng mục trong từng danh mục)

#Tạo biểu đồ tròn với phần trăm

ggplot(data1, aes(x = "", y = Frequency, fill = Menu\_Category)) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar("y", start=0) +

labs(title = "Phân phối mục theo danh mục thực đơn") +

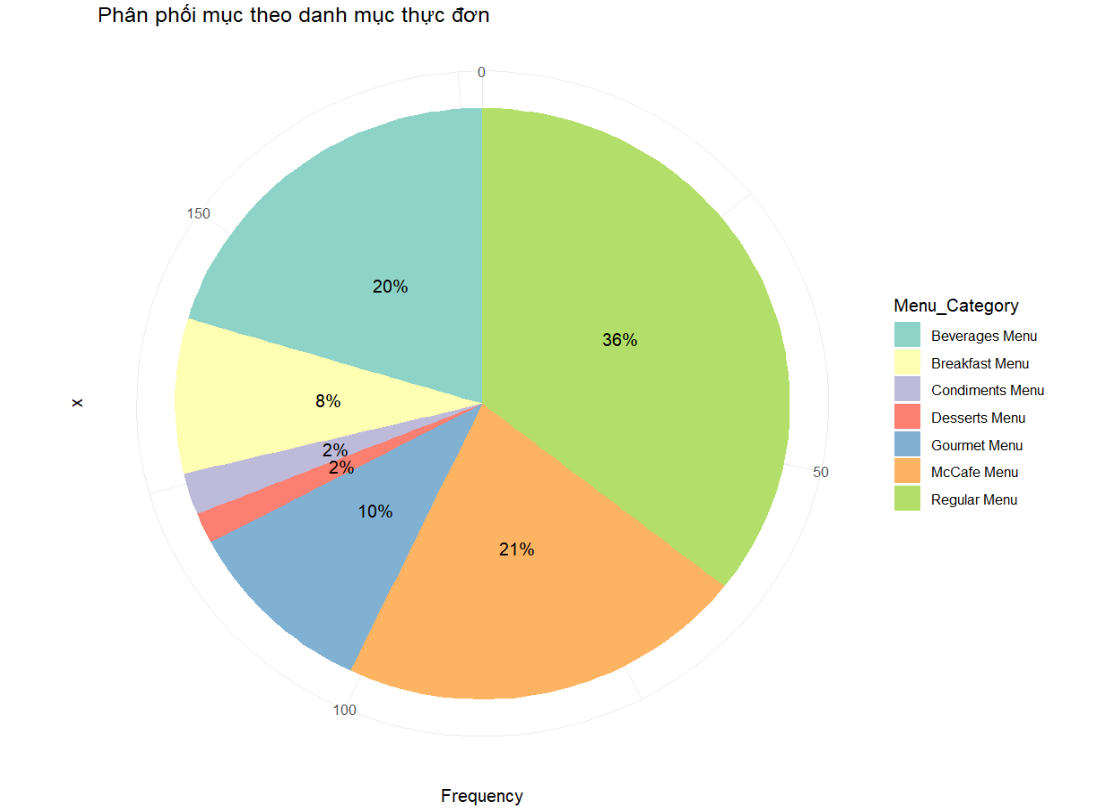
theme\_minimal() +

theme(legend.position = "right") +

scale\_fill\_brewer(palette = "Set3") + # Lựa chọn màu sắc từ bảng màu Set3

geom\_text(aes(label = paste0(round(Frequency / sum(Frequency) \* 100), "%")),

position = position\_stack(vjust = 0.5))



*Hình 3.1: Biểu đồ phân phối mục theo danh sách thực đơn*

* Biểu đồ tròn thể hiện phân phối mẫu dựa trên các danh mục thực đơn.
* Regular Menu: chiếm tỷ lệ lớn nhất trong biểu đồ với 36%, cho thấy đây là mặt hàng bán chạy nhất trong thực đơn.Các mặt hàng khác tương đối cao cho thấy sự đa dạng của thực đơn.Tuy nhiên,nhóm thực đơn Menu Gourmet và McCafe chiếm tỉ lệ thấp(8% và 3%) có thể do giá cao và ít phổ biến hơn.
* Biểu đồ mang lại lợi ích cho việc phân tích doanh số bán hàng,sở thích của khách hàng,hoặc quản lý dịch vụ tồn kho trong ngành dịch vụ ăn uống khách quan hơn.

### **3.2.2. Biểu đồ Bar chart**

**3.2.2.1. Biểu đồ bar chart của Energy**

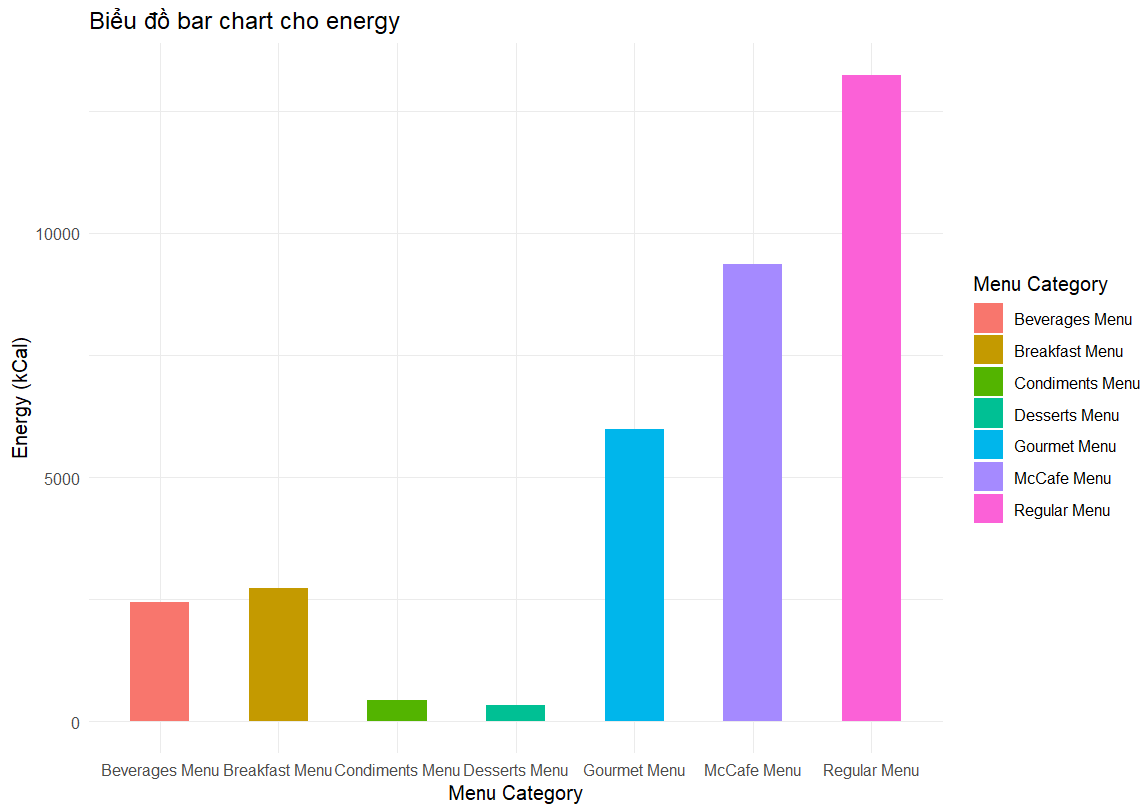
library(ggplot2)

ggplot(data = data, mapping = aes(y = `Energy (kCal)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho energy")+

theme\_minimal()



*Hình 3.2: Biểu đồ bar chart cho energy*

* Biểu đồ cho thấy sự phân bố năng lượng giữa các loại thực đơn. Mỗi danh mục thực đơn có một thanh cột tương ứng với lượng năng lượng trong danh mục đó.
* Regular Menu có lượng năng lượng cao nhất, gần 10,000 Kcal. Điều này có thể là do các món ăn trong thực đơn này có hàm lượng calo cao hơn so với các thực đơn khác. Các thực đơn khác như Beverages Menu, Condiments Menu, Desserts Menu và Gourmet Menu có lượng năng lượng thấp hơn nhiều so với các thực đơn trên.
* Biểu đồ này cho thấy sự khác biệt về lượng năng lượng giữa các loại thực đơn tại nhà hàng. Việc lựa chọn thực đơn phù hợp với nhu cầu năng lượng là quan trọng để duy trì sức khỏe và cân đối dinh dưỡng.

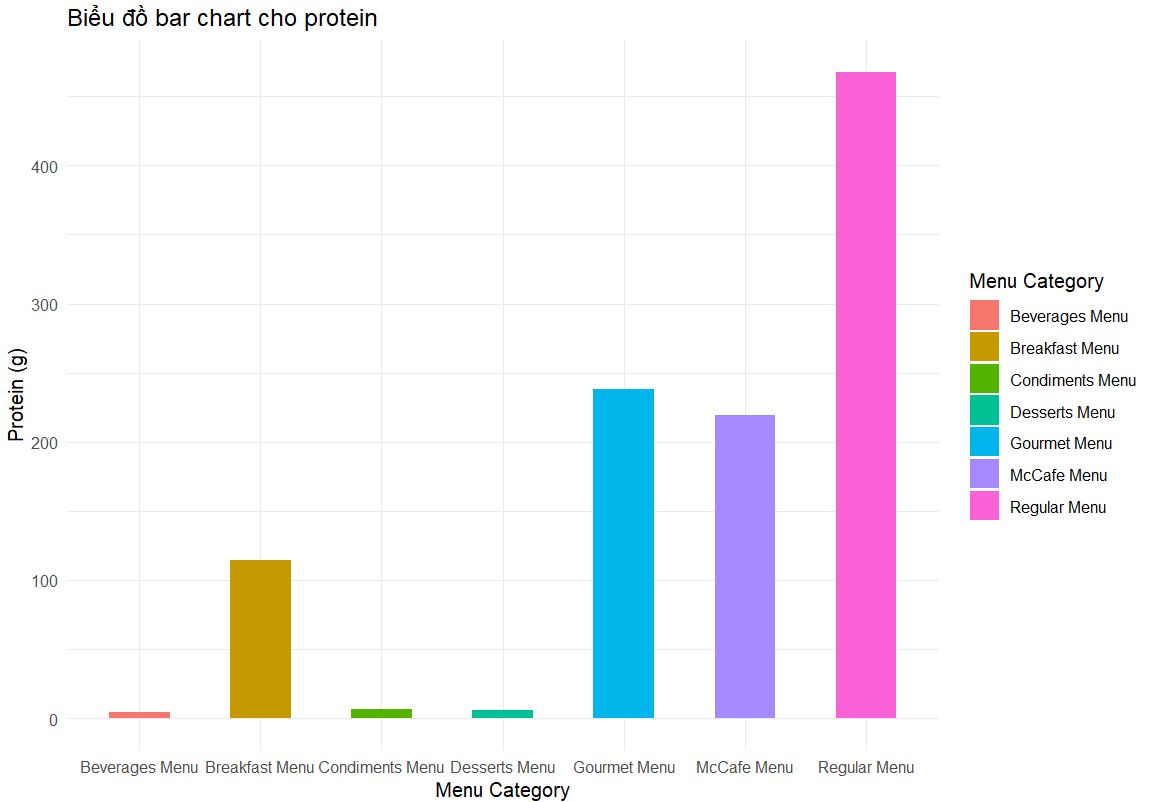
**3.2.2.2. Biểu đồ bar chart của Protein**

ggplot(data = data, mapping = aes(y =`Protein (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho protein")+

theme\_minimal()



*Hình 3.3: Biểu đồ bar chart cho protein*

* Biểu đồ này thể hiện nồng độ protein trong các danh mục thực đơn khác nhau.
* Có bảy thanh biểu đồ, mỗi thanh đại diện cho một danh mục thực đơn khác nhau: Beverages Menu, Breakfast Menu, Condiments Menu, Desserts Menu, Gourmet Menu, McCafe Menu, Regular Menu.
* Các thanh biểu đồ có chiều cao khác nhau, cho thấy lượng protein trong từng danh mục thực đơn.
* Regular Menu có nồng độ protein cao nhất được thể hiện bằng thanh màu hồng. Biểu đồ này cung cấp so sánh hình ảnh về nồng độ protein qua các loại thực đơn khác nhau, có thể hữu ích cho phân tích dinh dưỡng hoặc lập kế hoạch thực đơn.

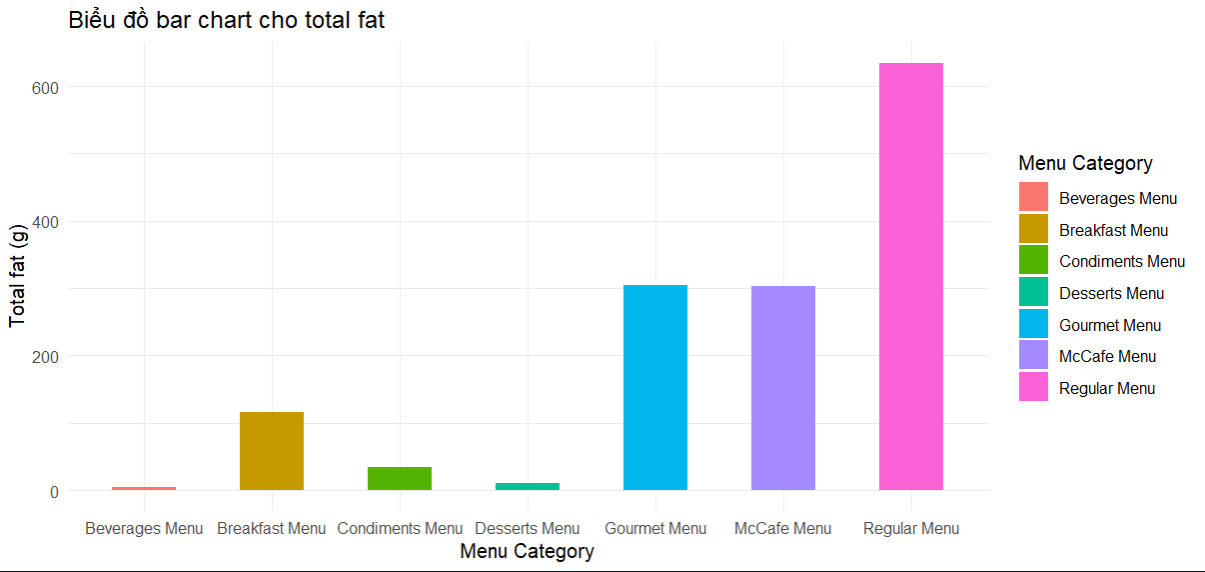
**3.2.2.3. Biểu đồ bar chart của Total fat**

ggplot(data = data, mapping = aes(y = `Total fat (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho total fat")+

theme\_minimal()



*Hình 3.4: Biểu đồ bar chart cho total fat*

* Biểu đồ này thể hiện nồng độ chất béo tổng hợp trong các danh mục thực đơn khác nhau.
* Regular Menu có nồng độ chất béo tổng hợp cao nhất với giá trị gần 600 gram, đáng kể hơn so với các danh mục khác.

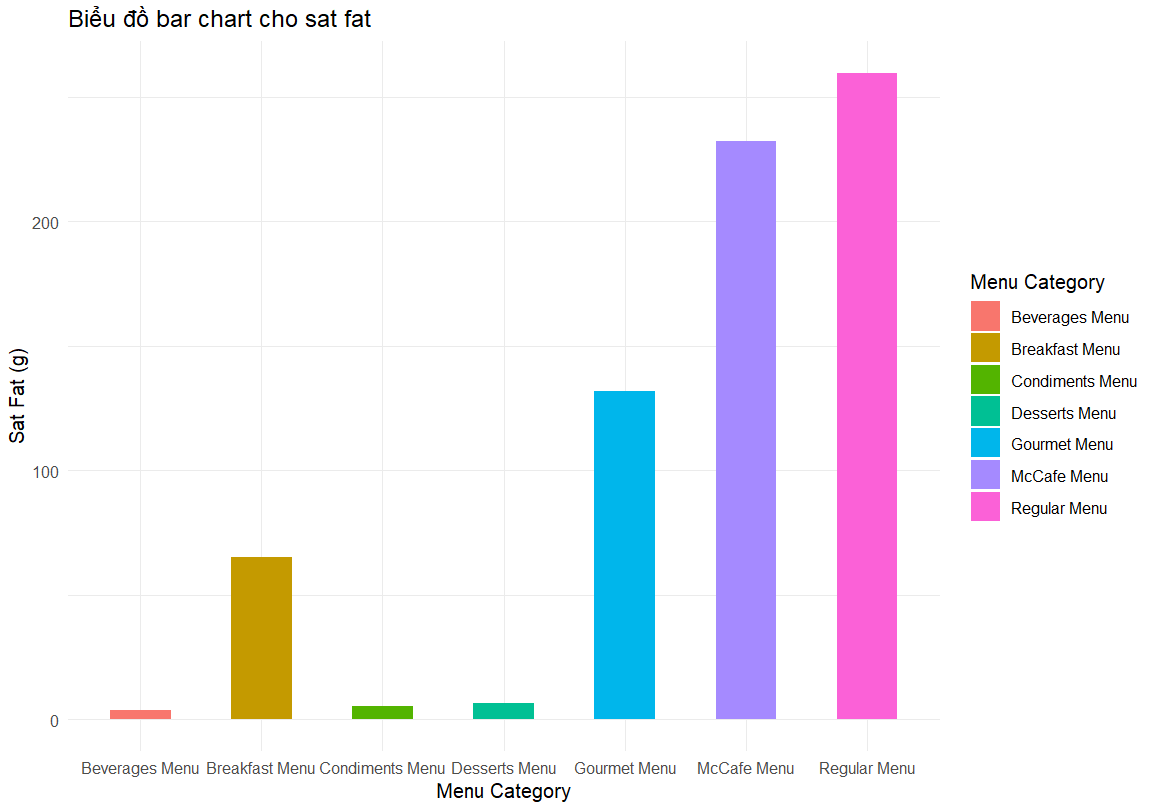
**3.2.2.4. Biểu đồ bar chart của Sat Fat**

ggplot(data = data, mapping = aes(y = `Sat Fat (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho sat fat")+

theme\_minimal()



*Hình 3.5: Biểu đồ bar chart cho sat fat*

* Biểu đồ này thể hiện lượng chất béo bão hòa (saturated fat) trong các danh mục thực đơn khác nhau.
* Desserts Menu: Đây là phần có lượng chất béo bão hòa cao nhất, khoảng 150-200 gram. Có lẽ vì các món tráng miệng thường chứa nhiều đường và bơ.
* Beverages Menu: Thực đơn này có lượng chất béo bão hòa thấp nhất, chỉ khoảng 10-15 gram.

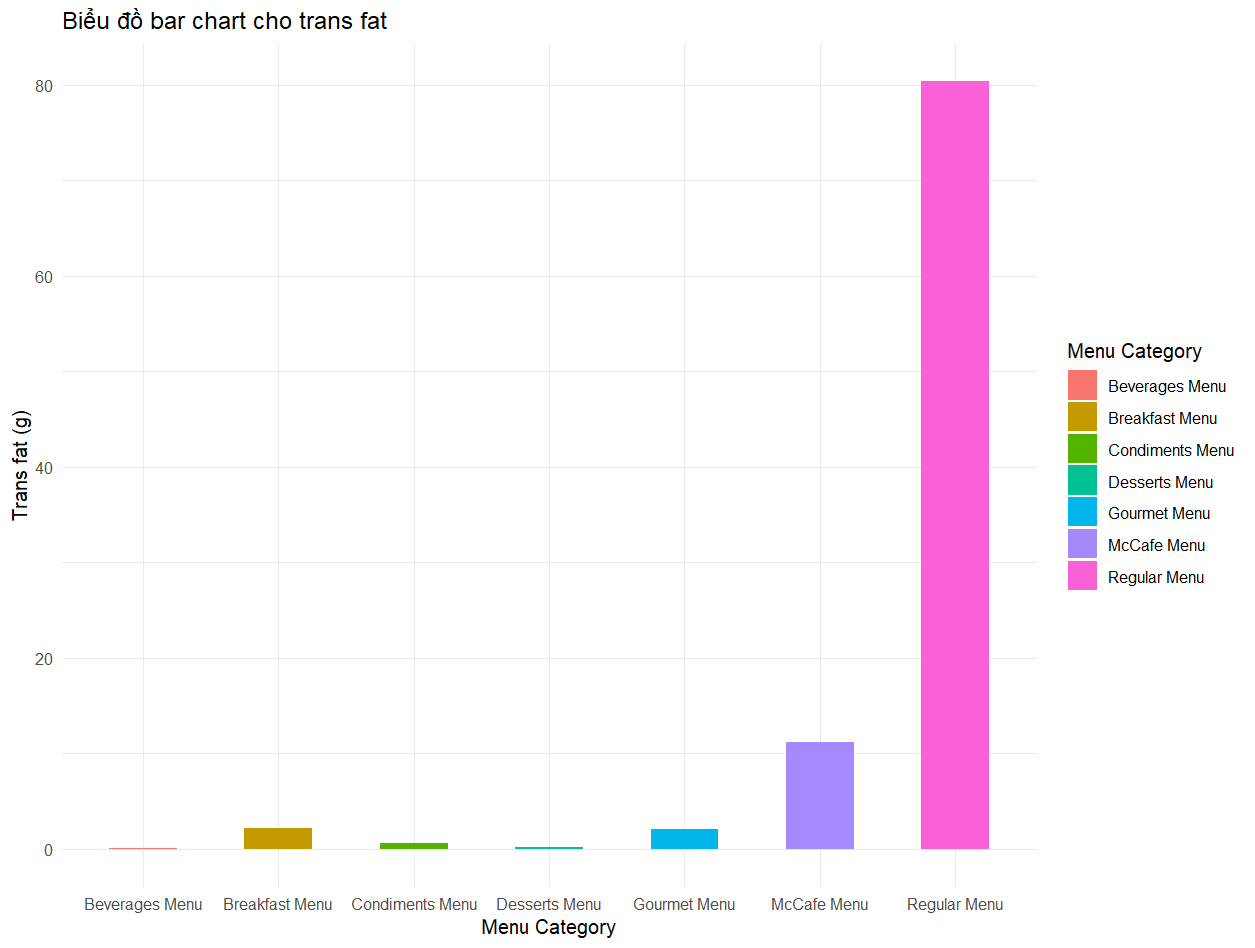
**3.2.2.5. Biểu đồ bar chart của Trans fat**

ggplot(data = data, mapping = aes(y = `Trans fat (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho trans fat")+

theme\_minimal()



*Hình 3.6: Biểu đồ bar chart cho Trans Fat*

* Biểu đồ này thể hiện lượng chất béo bão hòa (trans fat) trong các danh mục thực đơn khác nhau.
* Thực đơn Regular (Thực đơn thông thường) có lượng trans fat cao nhất, với giá trị khoảng 80g. Đây là mức trans fat cao hơn rất nhiều so với các danh mục khác.
* Thực đơn McCafe có lượng trans fat thấp hơn đáng kể, vào khoảng 10g.
* Các danh mục khác như Beverages, Breakfast, Condiments, Desserts và Gourmet có lượng trans fat không đáng kể, gần như bằng 0 hoặc chỉ vài gram.

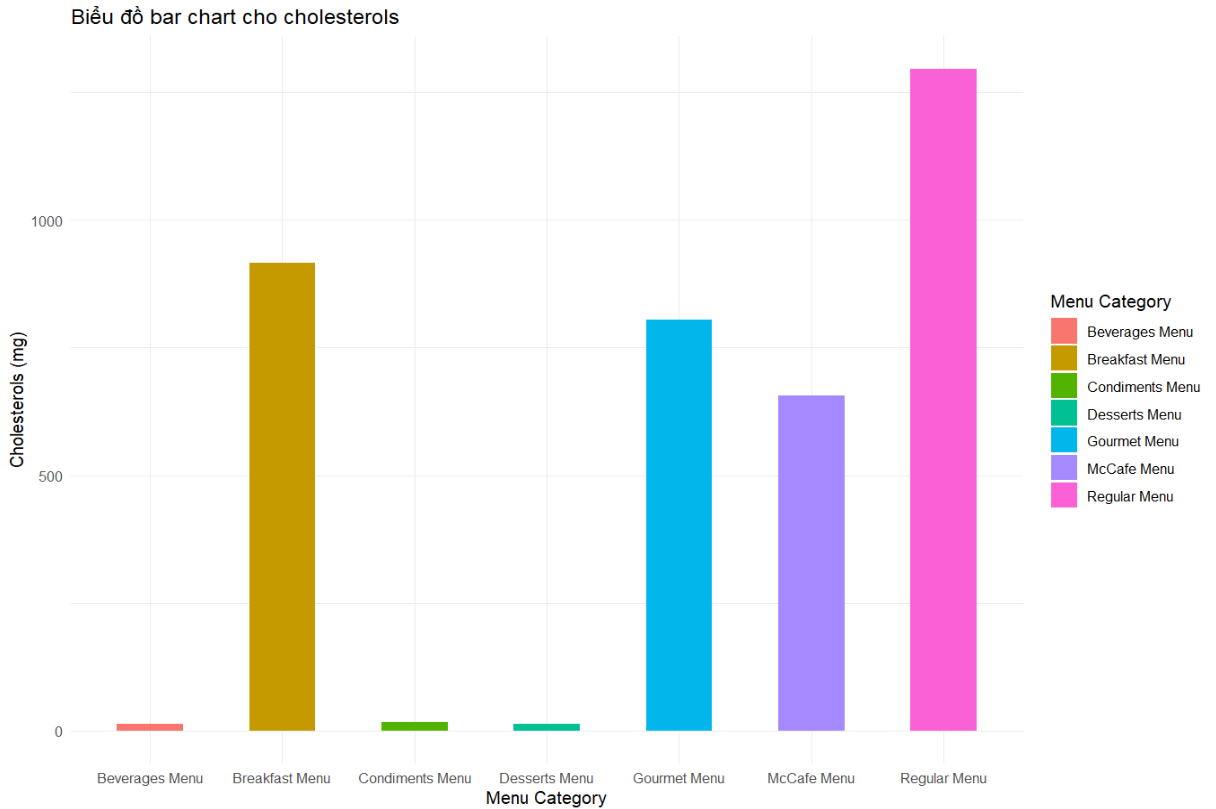
**3.2.2.6. Biểu đồ bar chart của Cholesterols**

ggplot(data = data, mapping = aes(y = `Cholesterols (mg)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho cholesterols")+

theme\_minimal()

****

*Hình 3.7: Biểu đồ bar chart cho cholesterols*

* Biểu đồ bar chart cho cholesterols (cholesterol) này thể hiện lượng cholesterol (mg) trong các danh mục thực đơn khác nhau.
* Thực đơn thông thường (Regular Menu) có lượng cholesterol cao nhất, cho thấy các món ăn trong danh mục này có hàm lượng cholesterol rất cao, có thể ảnh hưởng xấu đến sức khỏe tim mạch nếu tiêu thụ nhiều.
* Thực đơn bữa sáng (Breakfast Menu) và thực đơn sang trọng (Gourmet Menu) cũng có lượng cholesterol khá cao, cho thấy các món ăn trong những danh mục này cũng cần được tiêu thụ có kiểm soát.
* Các danh mục khác như Beverages, Condiments, và Desserts có lượng cholesterol rất thấp hoặc không đáng kể, ít gây lo ngại về vấn đề cholesterol.

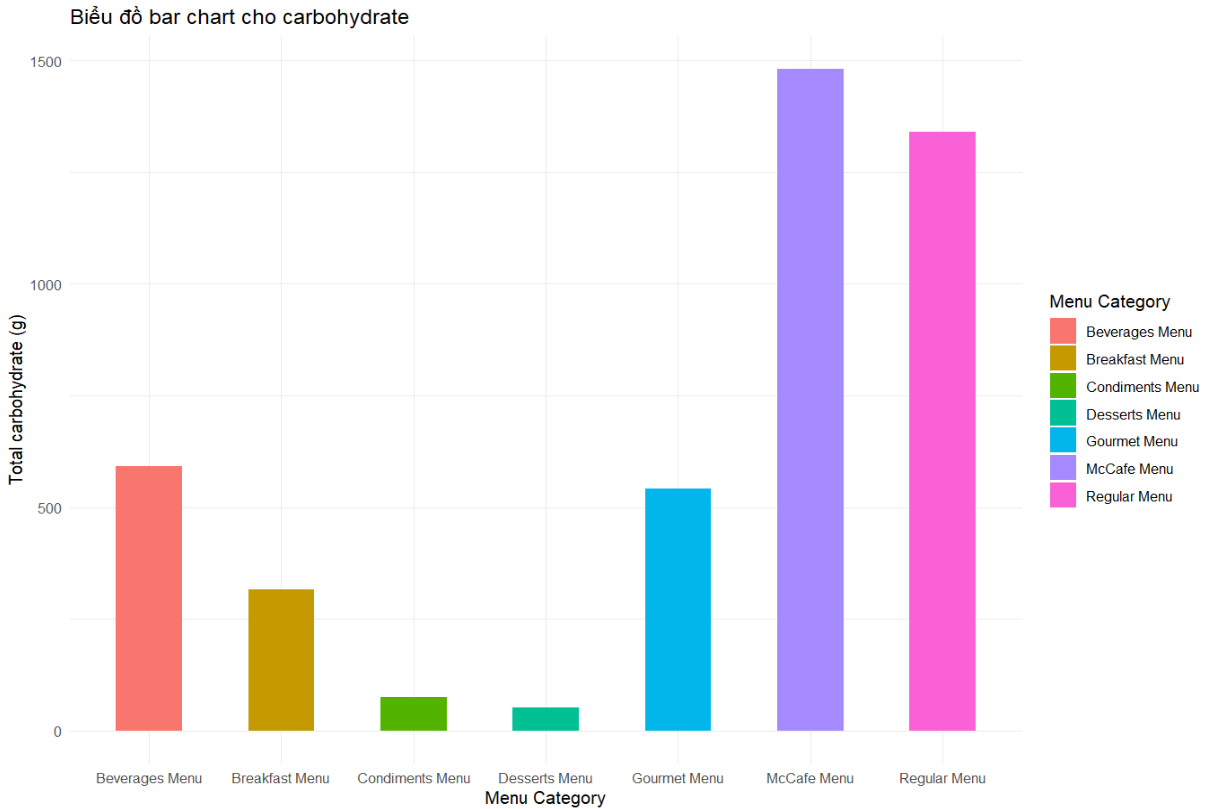
**3.2.2.7. Biểu đồ bar chart của Total carbohydrate**

ggplot(data = data, mapping = aes(y = `Total carbohydrate (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho carbohydrate")+

theme\_minimal()



*Hình 3.8: Biểu đồ bar chart cho total carbohydrate*

* Biểu đồ bar chart cho carbohydrate này thể hiện lượng tổng carbohydrate (g) trong các danh mục thực đơn khác nhau.
* Thực đơn McCafe có lượng carbohydrate cao nhất, cho thấy các món ăn trong danh mục này có hàm lượng carbohydrate rất cao, có thể ảnh hưởng đến mức đường huyết nếu tiêu thụ nhiều.
* Thực đơn thông thường (Regular Menu) cũng có lượng carbohydrate cao, cần được tiêu thụ có kiểm soát.
* Thực đơn đồ uống (Beverages Menu) và thực đơn bữa sáng (Breakfast Menu) có lượng carbohydrate tương đối, nhưng vẫn thấp hơn nhiều so với McCafe và Regular Menu.
* Các danh mục còn lại như Condiments và Desserts có lượng carbohydrate rất thấp hoặc không đáng kể, ít gây lo ngại về vấn đề carbohydrate.

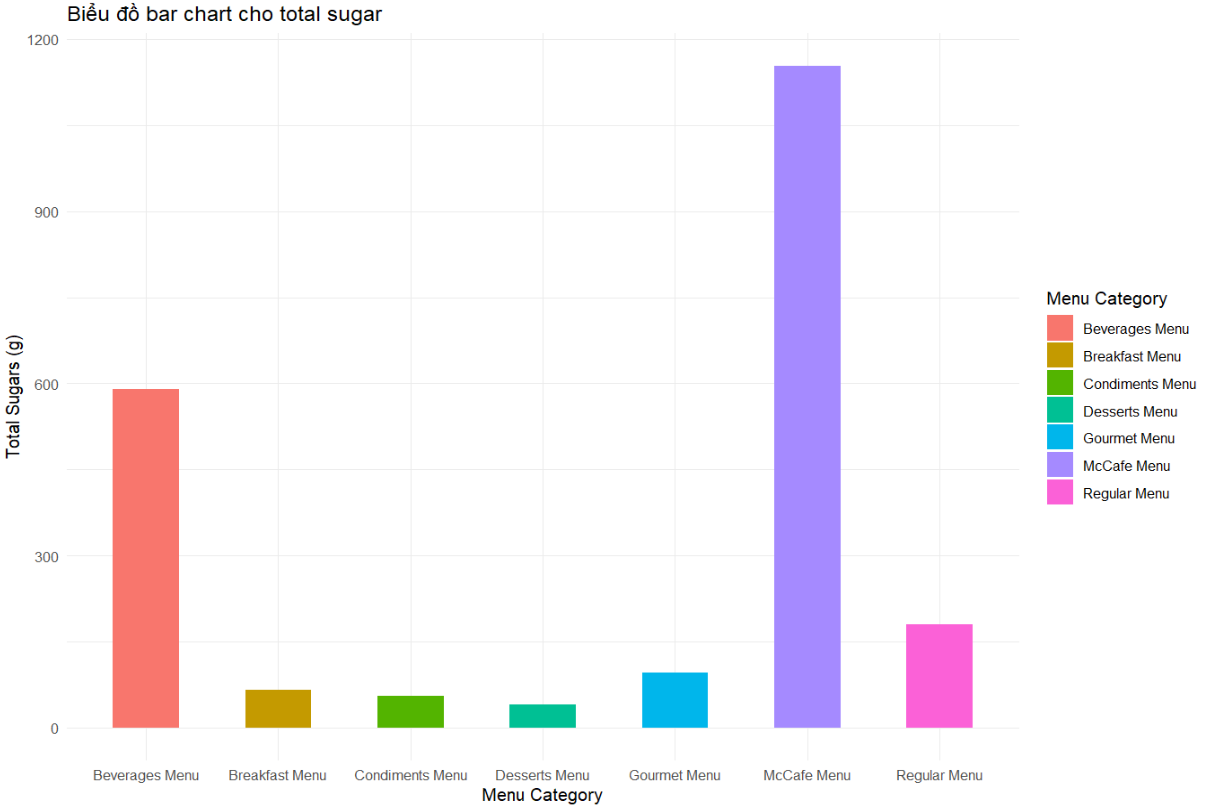
**3.2.2.8. Biểu đồ bar chart của Total Sugars**

ggplot(data = data, mapping = aes(y = `Total Sugars (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho total sugar")+

theme\_minimal()



*Hình 3.9: Biểu đồ bar chart cho total sugar*

* Biểu đồ trên biểu diễn lượng đường tổng cộng (total sugar) của các danh mục thực đơn (menu category) khác nhau.
* McCafe Menu và Beverages Menu là hai danh mục có lượng đường tổng cộng cao nhất, trong đó McCafe Menu chiếm ưu thế vượt trội.
* Breakfast Menu, Condiments Menu, Desserts Menu, và Gourmet Menu đều có lượng đường tổng cộng rất thấp, gần như không đáng kể so với McCafe Menu và Beverages Menu.
* Regular Menu có lượng đường tương đối cao nhưng vẫn thấp hơn nhiều so với hai danh mục đứng đầu.

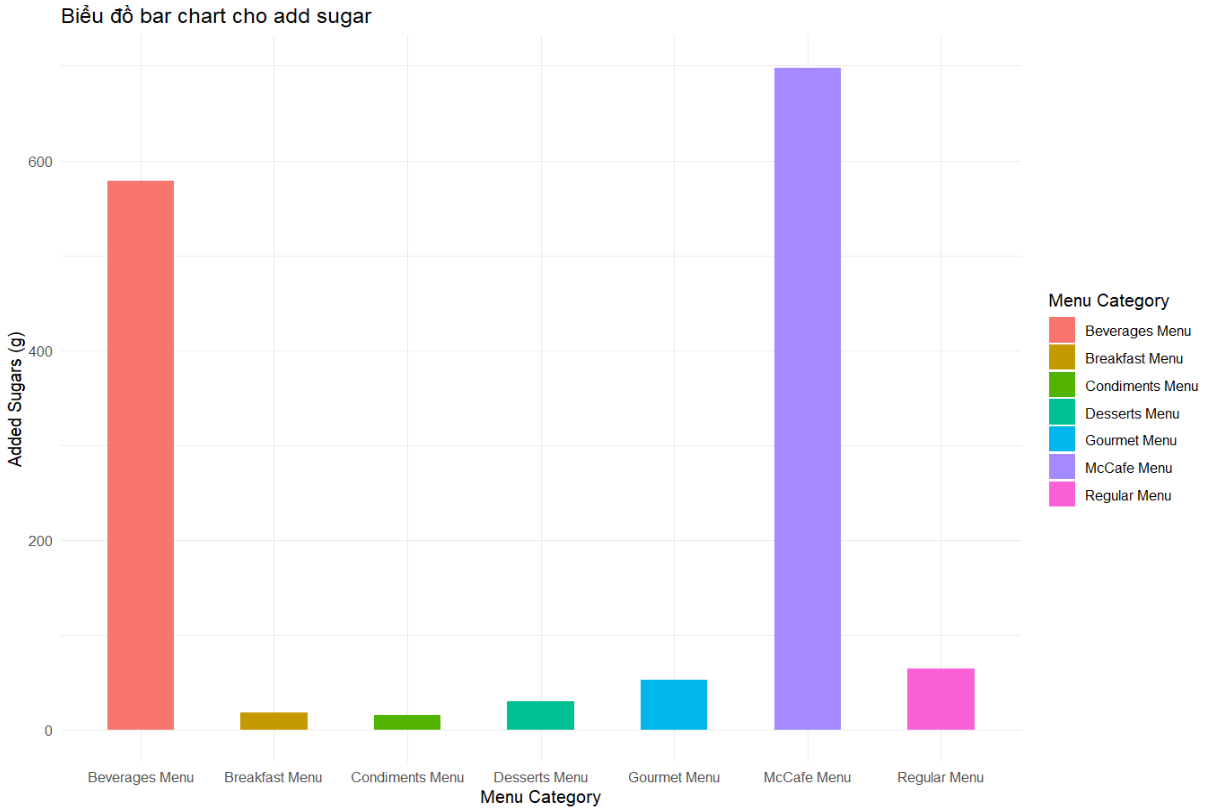
**3.2.2.9. Biểu đồ bar chart cho Added sugars**

ggplot(data = data, mapping = aes(y = `Added Sugars (g)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho add sugar")+

theme\_minimal()



*Hình 3.10: Biểu đồ bar chart cho added sugar*

* Biểu đồ trên biểu diễn lượng đường thêm vào (added sugar) của các danh mục thực đơn (menu category) khác nhau.
* McCafe Menu và Beverages Menu là hai danh mục có lượng đường thêm vào cao nhất, với McCafe Menu chiếm ưu thế nhẹ hơn so với Beverages Menu.
* Breakfast Menu, Condiments Menu, Desserts Menu, và Gourmet Menu đều có lượng đường thêm vào rất thấp, gần như không đáng kể so với McCafe Menu và Beverages Menu.
* Regular Menu có lượng đường thêm vào tương đối cao nhưng vẫn thấp hơn nhiều so với hai danh mục đứng đầu.

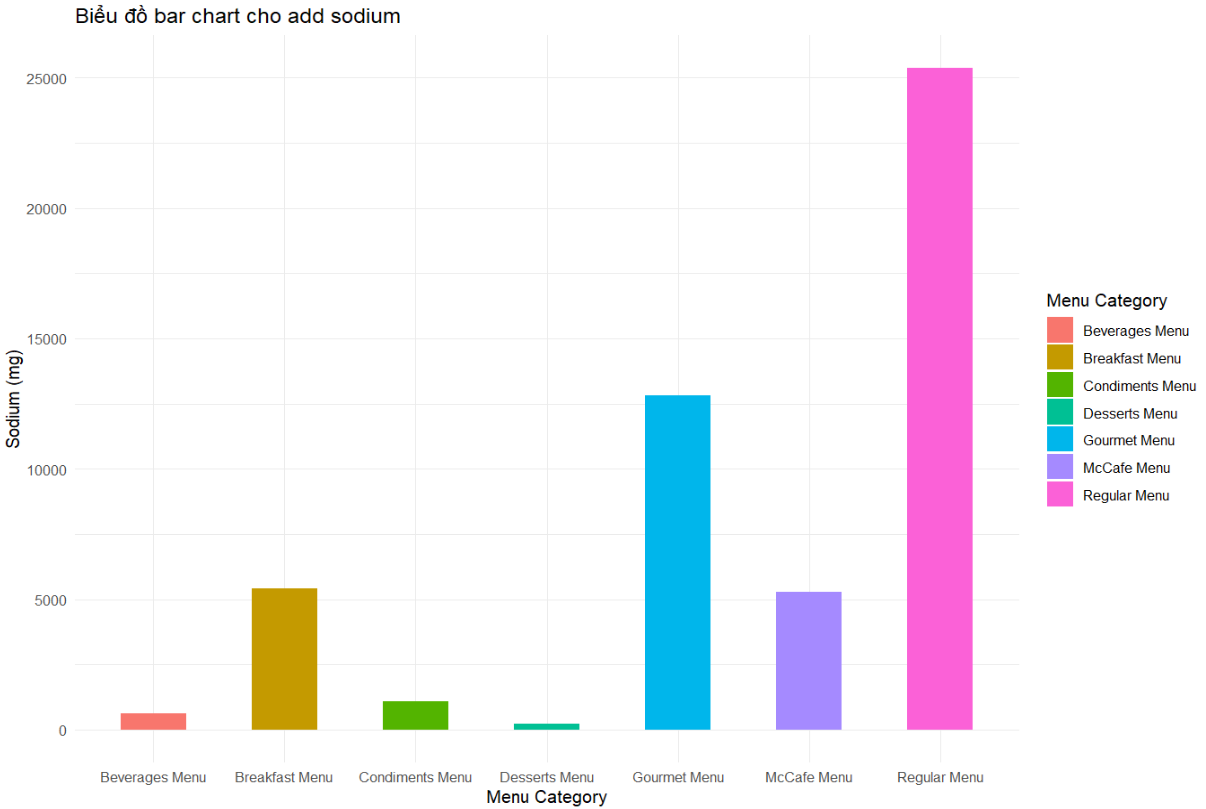
**3.2.2.10. Biểu đồ bar chart cho Sodium**

ggplot(data = data, mapping = aes(y = `Sodium (mg)`, x = `Menu Category`, fill = `Menu Category` )) +

geom\_bar(stat="identity", width=0.5)+

labs(title = "Biểu đồ bar chart cho add sodium")+

theme\_minimal()



*Hình 3.11: Biểu đồ bar chart cho sodium*

* Biểu đồ trên biểu diễn lượng natri (sodium) thêm vào của các danh mục thực đơn (menu category) khác nhau.
* Regular Menu có lượng natri thêm vào cao nhất, gần như gấp đôi so với Gourmet Menu và gấp hơn ba lần so với McCafe Menu.
* Beverages Menu và Desserts Menu có lượng natri thêm vào rất thấp, gần như không đáng kể so với các danh mục khác.
* Gourmet Menu và Breakfast Menu có lượng natri thêm vào ở mức trung bình, nhưng vẫn thấp hơn nhiều so với Regular Menu.
* Condiments Menu có lượng natri thêm vào thấp, nhưng cao hơn Beverages Menu và Desserts Menu.

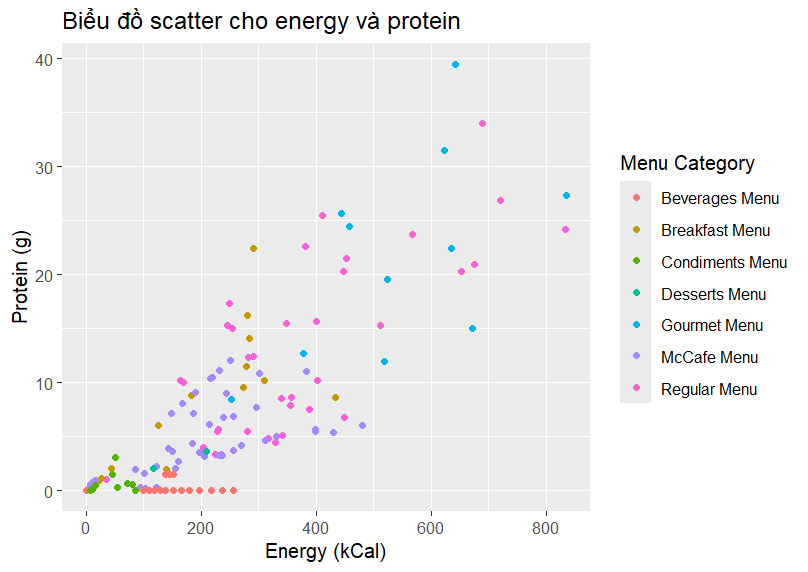
### **3.2.3. Biểu đồ Scatter plot**

**3.2.3.1. Biểu đồ scatter plot giữa Energy và Protein**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y =`Protein (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và protein")



*Hình 3.12: Biểu đồ scatter giữa energy và protein*

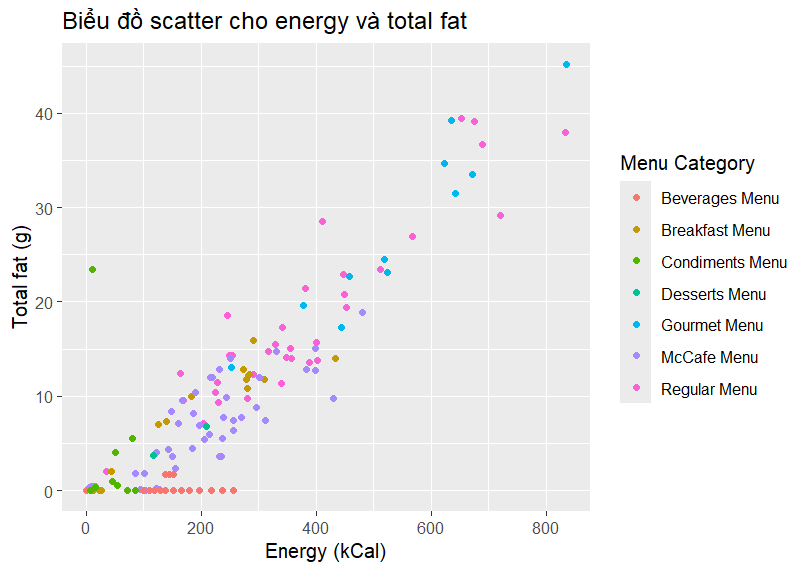
* Biểu đồ scatter này thể hiện mối quan hệ giữa năng lượng (đo bằng KCal) và hàm lượng protein (đo bằng g) trong danh mục thực đơn.
* Dữ liệu được hiển thị dưới dạng tập hợp các điểm, mỗi điểm có giá trị của một biến xác định vị trí trên trục hoành và giá trị của biến khác xác định vị trí trên trục tung.
* Năng lượng (Energy) và protein có một mức độ tương quan thấp đến trung bình. Điều này có nghĩa là khi năng lượng tăng, protein cũng có xu hướng tăng, nhưng không phải lúc nào cũng theo cùng một hướng.
* Hầu hết các điểm dữ liệu nằm trong khoảng từ 0 đến 400 KCal về năng lượng và từ 0 đến 10g về protein.
* Có một số điểm ngoại lai với giá trị năng lượng hoặc protein cao hơn.

**3.2.3.2. Biểu đồ scatter giữa Energy và Total fat**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Total fat (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và total fat")



*Hình 3.13: Biểu đồ scatter giữa energy và total fat*

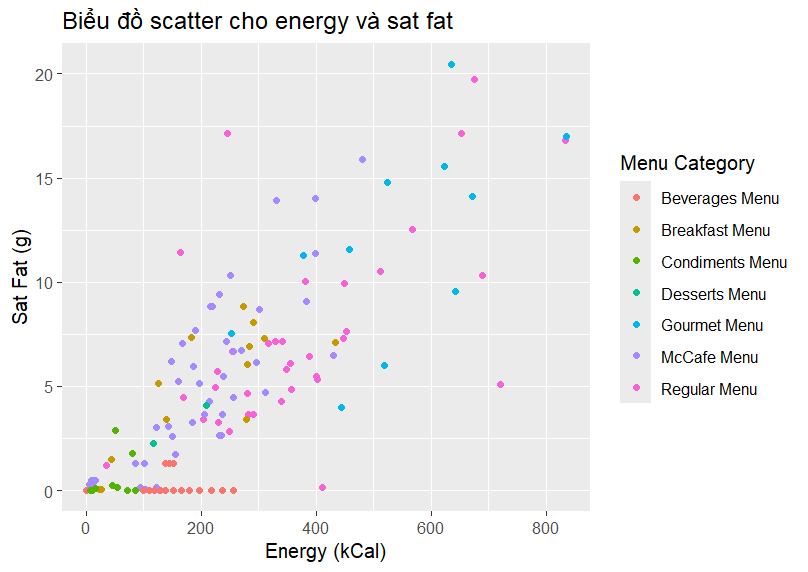
* Xu hướng Tương quan: Có một xu hướng tương quan dương giữa năng lượng và tổng lượng mỡ, nghĩa là khi năng lượng tăng thì tổng lượng mỡ cũng tăng theo.
* Phân bố Dữ liệu: Các điểm dữ liệu phân bố chủ yếu theo hướng tăng từ trái sang phải trên biểu đồ, cho thấy mối quan hệ này.
* Biểu đồ này hữu ích trong việc phân tích dinh dưỡng hoặc lựa chọn thực phẩm dựa trên lượng calo và mỡ tiêu thụ.

**3.2.3.3. Biểu đồ scatter giữa Energy và Sat fat**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Sat Fat (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và sat fat")



*Hình 3.14: Biểu đồ scatter giữa energy và sat fat*

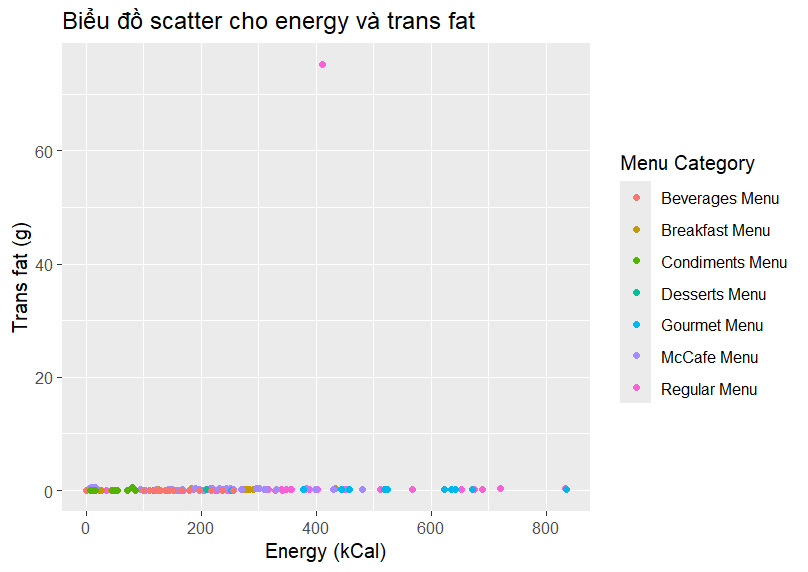
* Có một mối quan hệ dương giữa năng lượng và chất béo bão hòa. Khi năng lượng tăng, chất béo bão hòa cũng có xu hướng tăng, cho thấy mối quan hệ tích cực giữa hai yếu tố này.
* Có một xu hướng rõ ràng: sự tăng năng lượng thường đi kèm với sự tăng chất béo bão hòa trong tất cả các danh mục thực đơn.
* Biểu đồ này có thể hữu ích cho việc phân tích dinh dưỡng hoặc lập kế hoạch ăn uống, giúp người dùng hiểu rõ hơn về mối quan hệ giữa năng lượng và chất béo bão hòa trong thực phẩm.

**3.2.3.4. Biểu đồ scatter giữa Energy và Trans fat**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Trans fat (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và trans fat")



*Hình 3.15: Biểu đồ scatter giữa energy và trans fat*

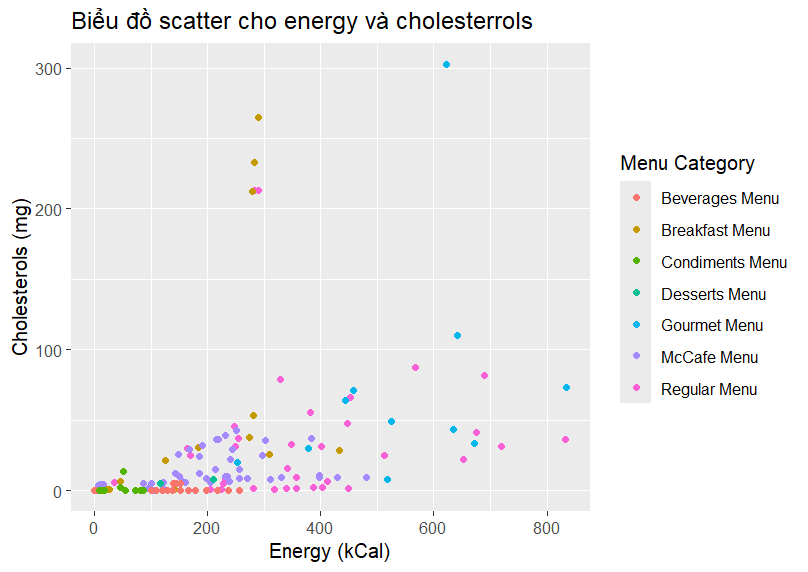
* Biểu đồ phân tán cho chúng ta biết mối quan hệ giữa các danh mục thực đơn và năng lượng, trans fat.
* Có một xu hướng tỉ lệ thuận giữa năng lượng (KCal) và trans fat (g). Nếu năng lượng tăng, trans fat cũng có xu hướng tăng theo.
* Tuy nhiên, chúng ta cũng thấy rằng có một số mục có năng lượng cao nhưng vẫn giữ trans fat ở mức thấp.

**3.2.3.5. Biểu đồ scatter giữa Energy và Cholesterol**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Cholesterols (mg)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và cholesterol")



*Hình 3.16: Biểu đồ scatter giữa energy và cholesterols*

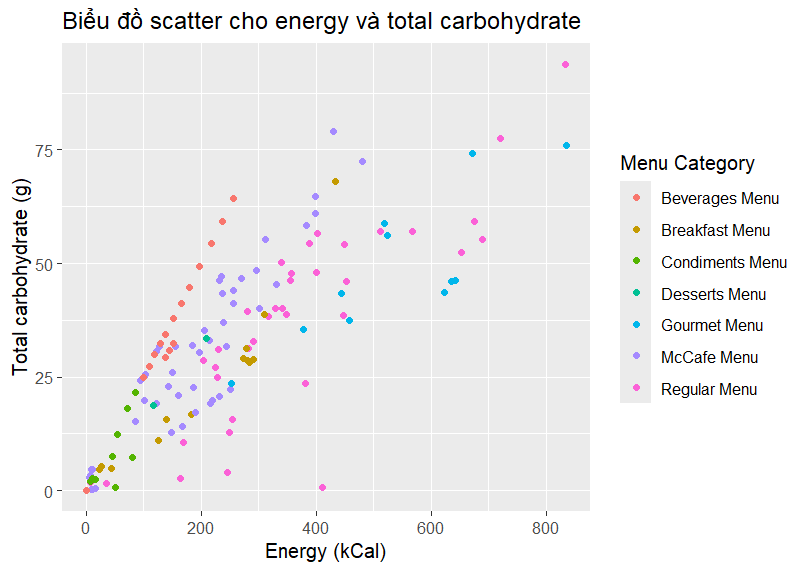
* Biểu đồ cho thấy sự phân tán rộng lớn của các điểm dữ liệu, không có một mẫu tương quan rõ ràng giữa năng lượng và cholesterol. Điều này cho thấy có những yếu tố khác ảnh hưởng đến hàm lượng cholesterol ngoài năng lượng.
* Không có xu hướng tuyến tính rõ ràng giữa năng lượng và cholesterol. Có thể có sự biến đổi lớn trong hàm lượng cholesterol ngay cả khi năng lượng tương tự.

**3.2.3.6. Biểu đồ scatter giữa Energy và Carbohydrate**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Total carbohydrate (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và total carbohydrate")



*Hình 3.17: Biểu đồ scatter giữa energy và total carbohydrate*

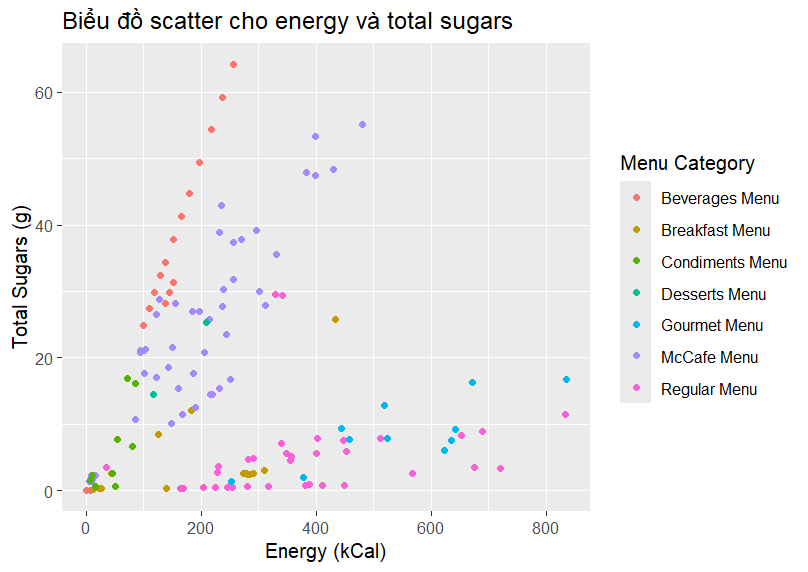
* Có một tương quan dương giữa năng lượng và lượng carbohydrate tổng cộng. Tức là khi năng lượng tăng, lượng carbohydrate cũng tăng. Tuy nhiên, tương quan này không rất mạnh vì có nhiều điểm ngoại lai và dữ liệu phân tán không theo một đường thẳng chính xác.
* Biểu đồ cho thấy rằng các loại thực đơn khác nhau có sự biến đổi về năng lượng và lượng carbohydrate.

**3.2.3.7. Biểu đồ scatter giữa Energy và Total sugars**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Total Sugars (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và total sugars")



*Hình 3.18: Biểu đồ scatter giữa energy và total sugars*

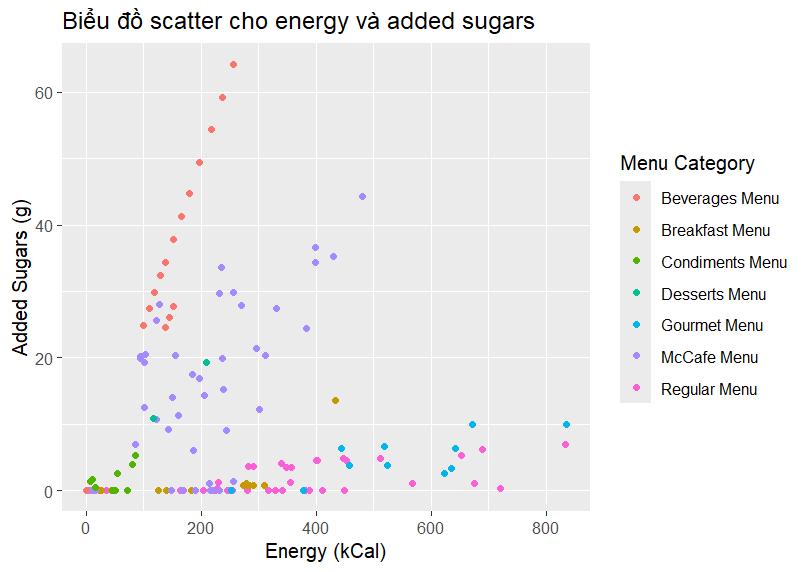
* Các điểm dữ liệu được phân bố theo các nhóm thực đơn khác nhau, cho thấy sự đa dạng về mức năng lượng và hàm lượng đường trong từng nhóm.
* Biểu đồ có thể hữu ích trong việc phân tích dinh dưỡng hoặc lập kế hoạch ăn uống, giúp người xem hiểu được mối liên hệ giữa lượng năng lượng tiêu thụ và lượng đường trong thực phẩm.

**3.2.3.8. Biểu đồ scatter giữa Energy và Added sugars**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Added Sugars (g)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và added sugars")



*Hình 3.19: Biểu đồ scatter giữa energy và added sugars*

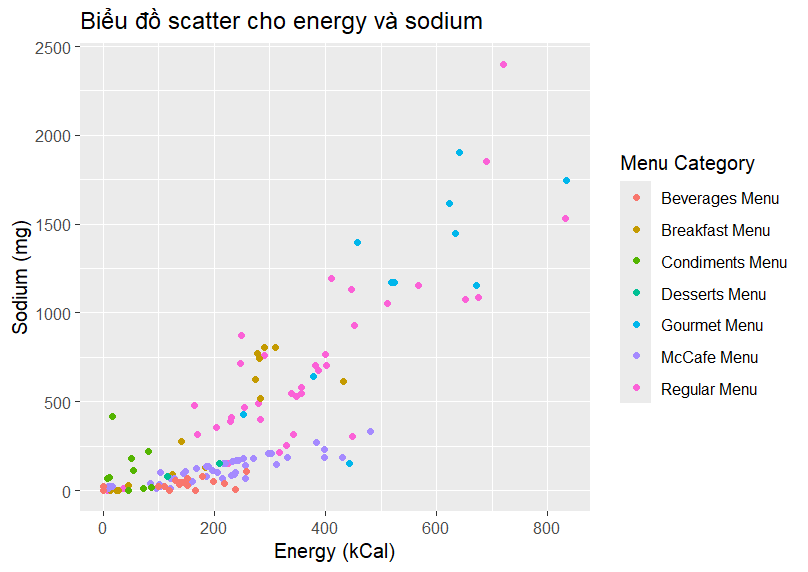
* Phân bố Dữ liệu: Hầu hết các điểm dữ liệu tập trung ở phía dưới cả hai trục, với một số điểm ngoại lệ ở giá trị cao hơn.
* Mối Quan Hệ: Không có mối quan hệ tuyến tính rõ ràng giữa năng lượng và đường thêm vào vì các điểm dữ liệu không tạo thành một đường thẳng mà phân tán trên biểu đồ.
* Biểu đồ này hữu ích trong việc phân tích dinh dưỡng hoặc lập kế hoạch ăn uống bằng cách cho thấy sự biến đổi giữa hàm lượng calo và đường trong các loại thực phẩm khác nhau.

**3.2.3.9. Biểu đồ scatter giữa Energy và Sodium**

ggplot(data = data, mapping = aes(x = `Energy (kCal)`, y = `Sodium (mg)`, color = `Menu Category`))+

geom\_point(alpha = 3)+

labs(title = "Biểu đồ scatter cho energy và sodium")



*Hình 3.20: Biểu đồ scatter giữa energy và sodium*

* Phân bố Dữ liệu: Có vẻ như có một nhóm các điểm dữ liệu với giá trị năng lượng thấp có lượng sodium thay đổi, trong khi các giá trị năng lượng cao hơn có xu hướng có lượng sodium cao hơn, cho thấy một mối quan hệ dương tiềm năng giữa lượng calo trong các món ăn và lượng sodium của chúng.
* Các món ăn từ danh mục ‘Regular Menu’ có xu hướng có lượng calo và sodium cao hơn, điều này có thể quan trọng đối với việc phân tích dinh dưỡng hoặc lập kế hoạch ăn uống.

### **3.2.4. Biểu đồ Boxplot**

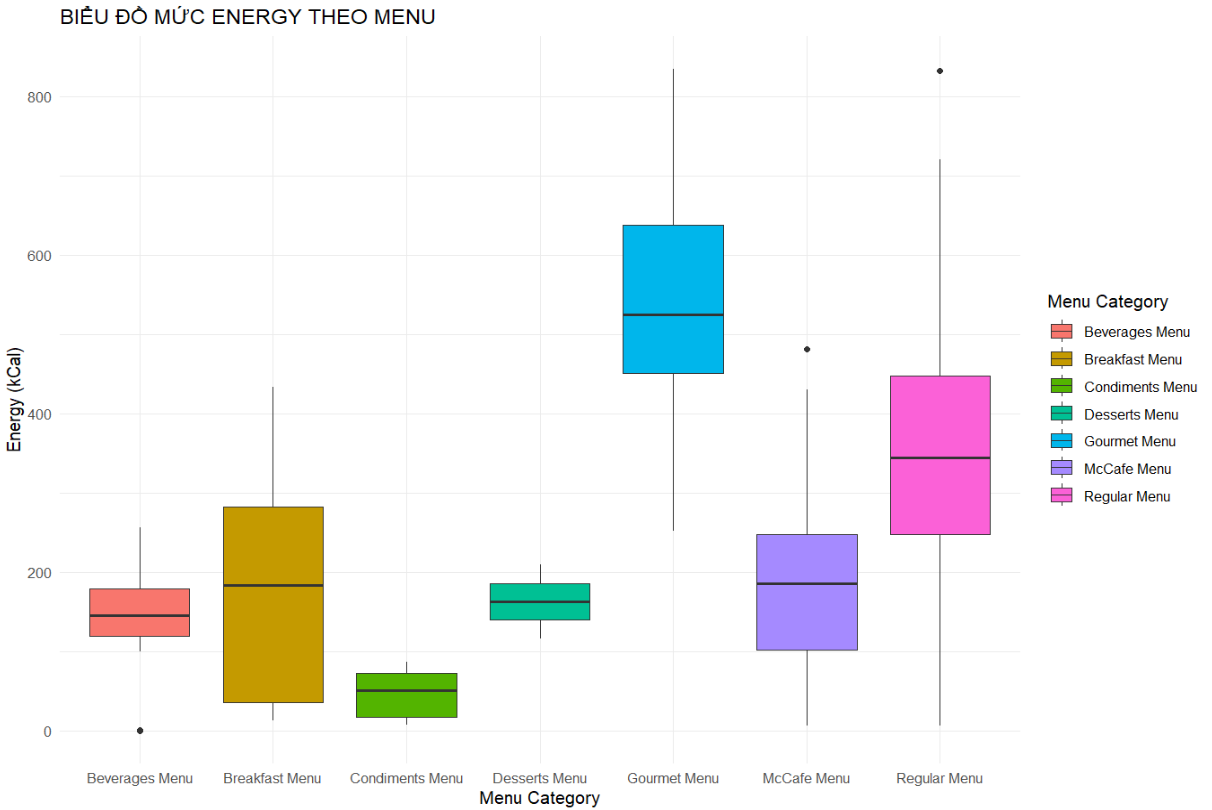
**3.2.4.1. Biểu đồ boxplot của Energy**

ggplot(data = data, mapping = aes(y = `Energy (kCal)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC ENERGY THEO MENU")



*Hình 3.21: Biểu đồ boxplot của energy*

* Biểu đồ thể hiện sự phân bố mức năng lượng (đo bằng kilocalories) qua các loại thực đơn khác nhau.
* Có sự đa dạng về mức năng lượng giữa các loại thực đơn từ Beverages đến Regular Menu.
* Khoảng Năng Lượng: Yếu tố này cho thấy sự chênh lệch lớn về năng lượng giữa các món ăn trong cùng một loại thực đơn.
* Biểu đồ này cung cấp cái nhìn trực quan về cách thức năng lượng được phân bố trong thực đơn,giúp người xem có thể đánh giá và lựa chọn món ăn phù hợp với nhu cầu năng lượng của bản thân.

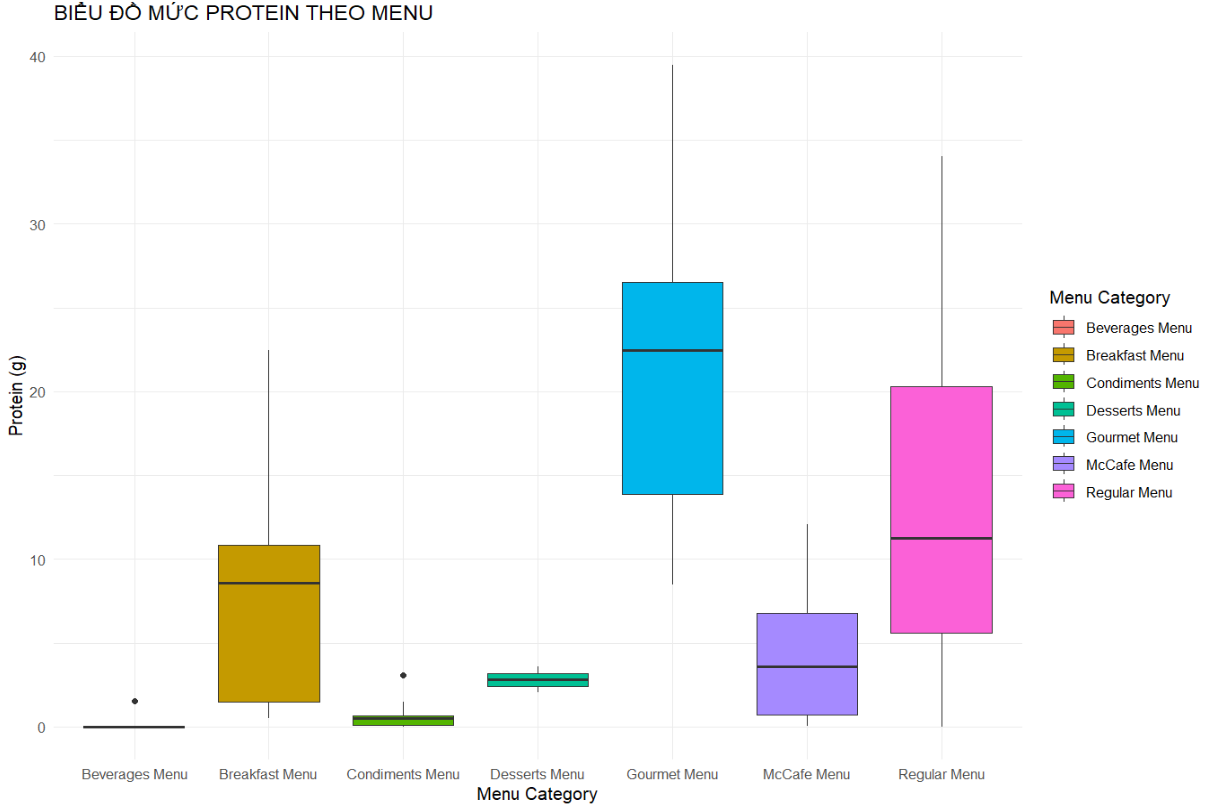
**3.2.4.2. Biểu đồ boxplot của Protein**

ggplot(data = data, mapping = aes(y = `Protein (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC PROTEIN THEO MENU")



*Hình 3.22: Biểu đồ boxplot của protein*

* Có sự đa dạng về hàm lượng protein giữa các loại thực đơn, thể hiện qua kích thước và vị trí của các box plot.
* Biểu đồ cung cấp một cách so sánh trực quan về nội dung protein, hữu ích cho việc phân tích dinh dưỡng hoặc lập kế hoạch thực đơn.

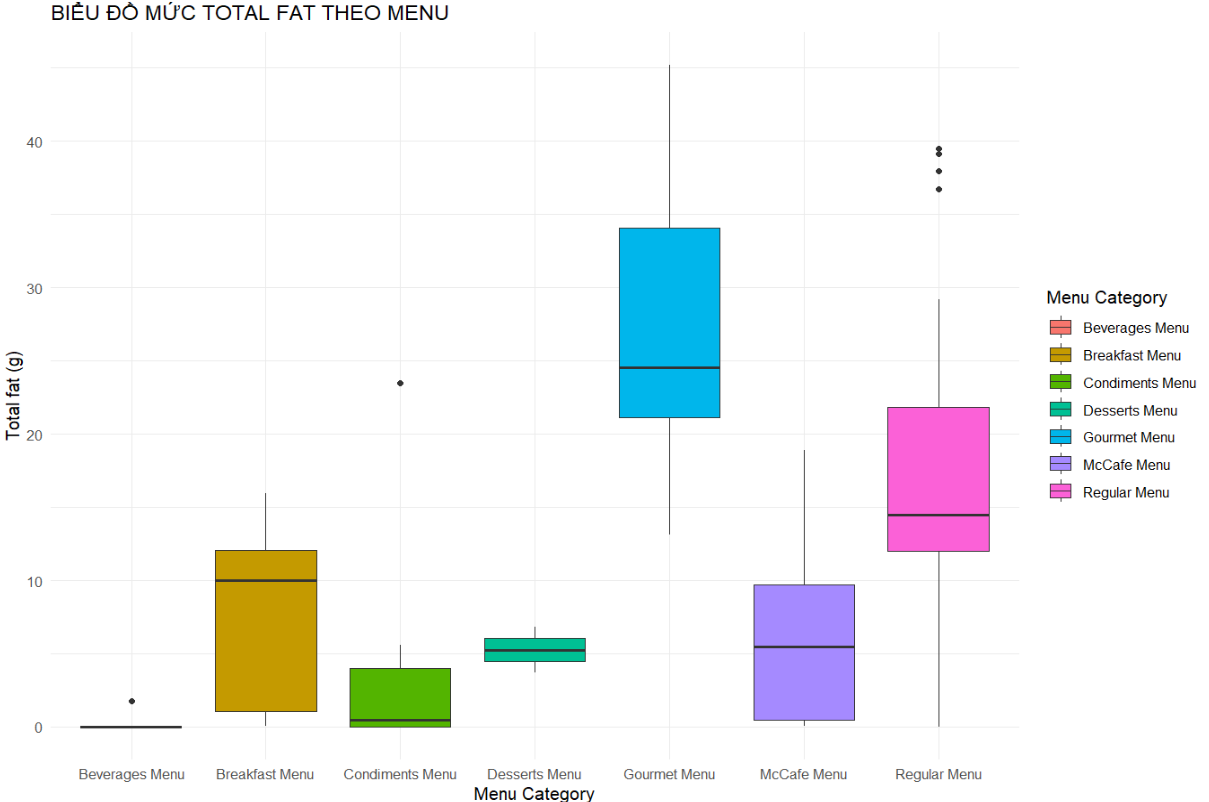
**3.2.4.3. Biểu đồ boxplot của Total fat**

ggplot(data = data, mapping = aes(y = `Total fat (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC TOTAL FAT THEO MENU")



*Hình 3.23: Biểu đồ boxplot của total fat*

* Sự biến động: Các đường kéo dài từ hộp (whiskers) chỉ ra sự biến động của lượng chất béo ngoài phạm vi tứ phân vị, và các điểm đơn lẻ biểu diễn các giá trị ngoại lệ.
* So sánh nhanh: Biểu đồ cung cấp một cái nhìn tổng quan để so sánh nhanh lượng chất béo giữa các loại thực đơn, có ích cho việc phân tích dinh dưỡng hoặc lập kế hoạch thực đơn.

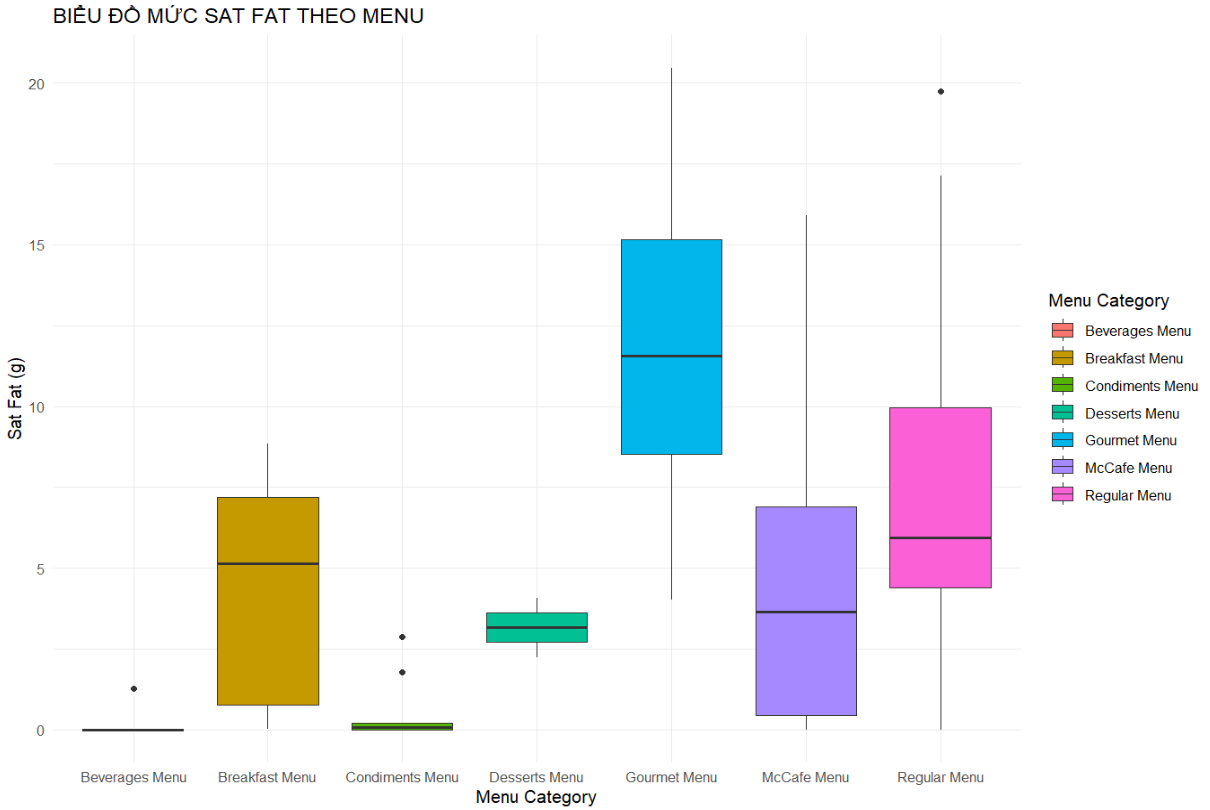
**3.2.4.4. Biểu đồ boxplot của Sat fat**

ggplot(data = data, mapping = aes(y = `Sat Fat (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC SAT FAT THEO MENU")



*Hình 3.24: Biểu đồ boxplot của sat fat*

* Biểu đồ trên thể hiện mức độ chất béo bão hòa (Sat Fat) theo từng danh mục thực đơn khác nhau.
* Thực đơn cao cấp (Gourmet Menu) có mức chất béo bão hòa cao nhất, trong khi Đồ uống (Beverages Menu) và Gia vị (Condiments Menu) có mức thấp nhất.
* Breakfast Menu và McCafe có mức độ chất béo bão hòa trung bình, còn Thực đơn tráng miệng và thông thường có mức độ biến đổi đa dạng hơn.

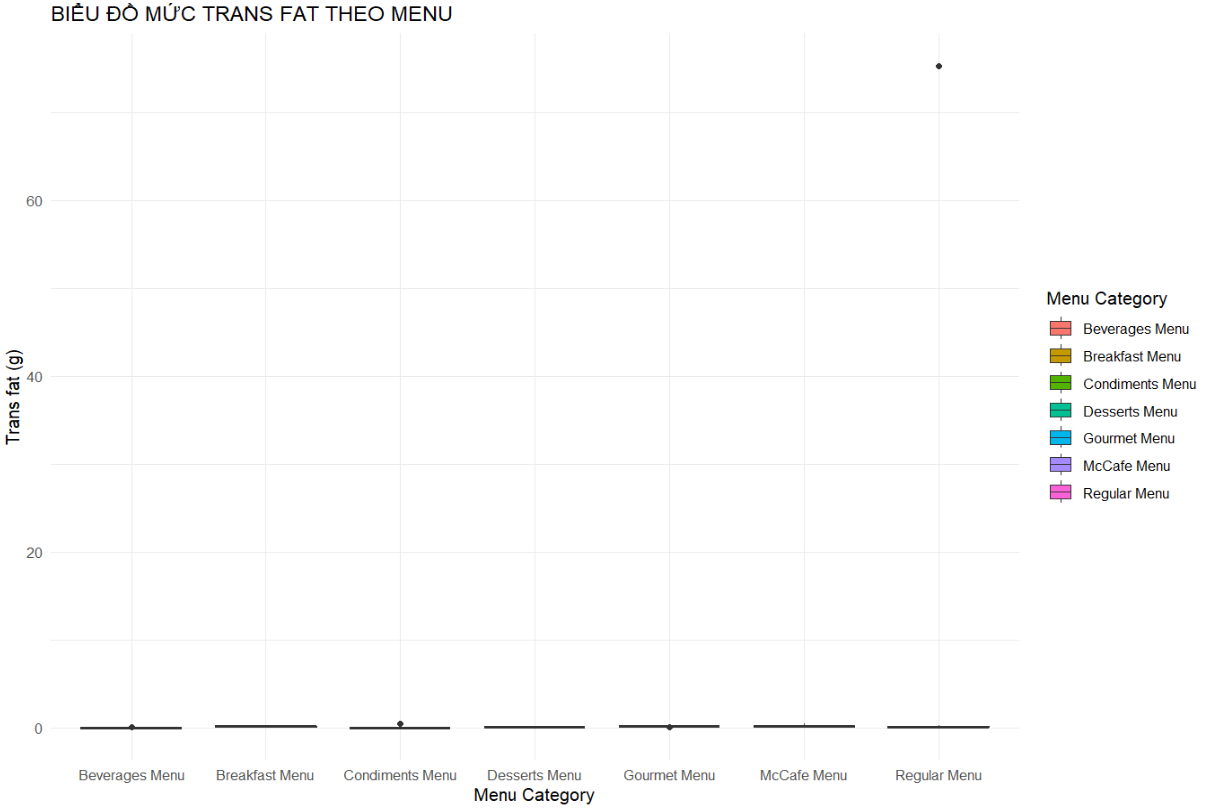
**3.2.4.5. Biểu đồ boxplot của Trans fat**

ggplot(data = data, mapping = aes(y = `Trans fat (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC TRANS FAT THEO MENU")



*Hình 3.25: Biểu đồ boxplot của trans fat*

* Hầu hết các danh mục menu đều có lượng chất béo trans rất thấp, gần như bằng 0 gram.
* Có một danh mục có giá trị rất cao so với các danh mục còn lại. Điểm này nằm ở danh mục Gourmet Menu với lượng chất béo trans trên 60 gram.

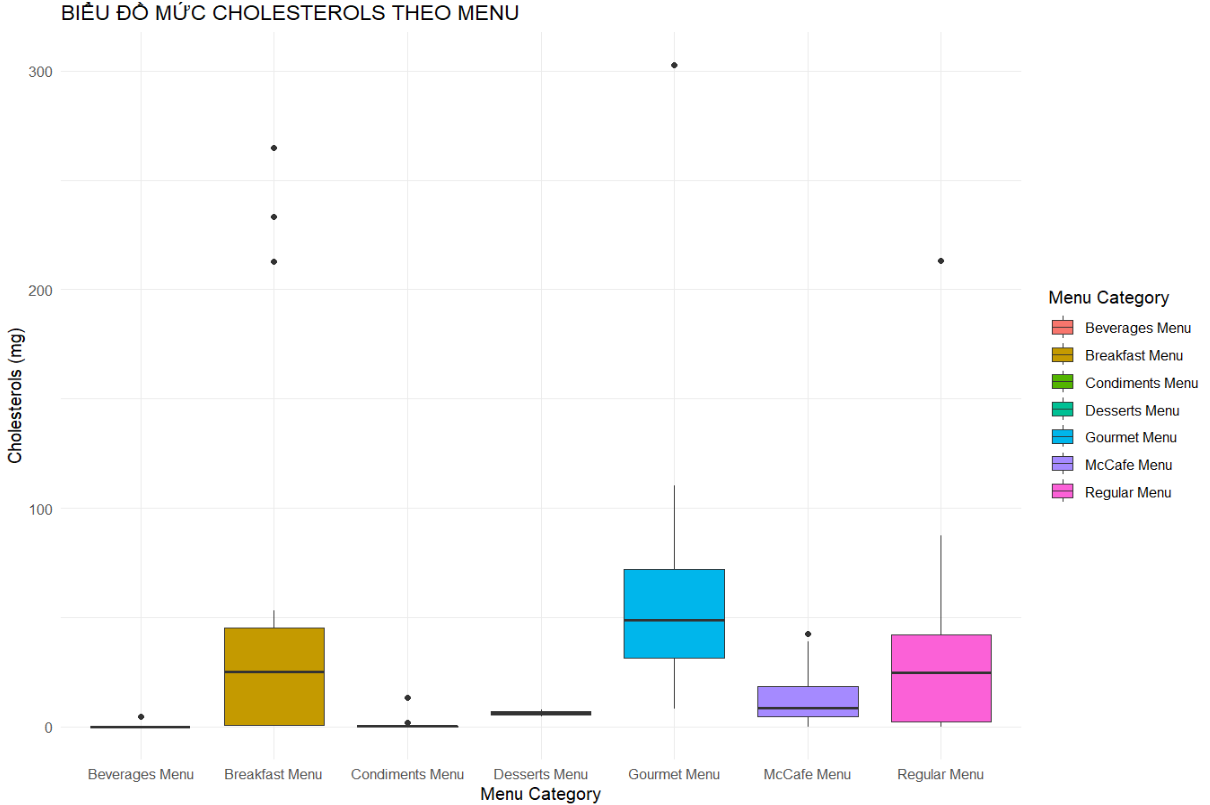
**3.2.4.6. Biểu đồ boxplot của Cholesterol**

ggplot(data = data, mapping = aes(y = `Cholesterols (mg)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC CHOLESTEROL THEO MENU")



*Hình 3.26: Biểu đồ boxplot của cholesterols*

* Biểu đồ này thể hiện mức cholesterol (mg) trong các danh mục menu khác nhau tại một nhà hàng.
* Các danh mục Breakfast, Desserts, Gourmet, và Regular Menu có mức cholesterol cao hơn so với các danh mục khác.
* Danh mục Beverages Menu và Condiments Menu có mức cholesterol rất thấp.
* Sự hiện diện của các giá trị ngoại lệ trong các danh mục Breakfast, Desserts, Gourmet, và Regular cho thấy rằng có một số món ăn trong các danh mục này có mức cholesterol rất cao so với phần còn lại.

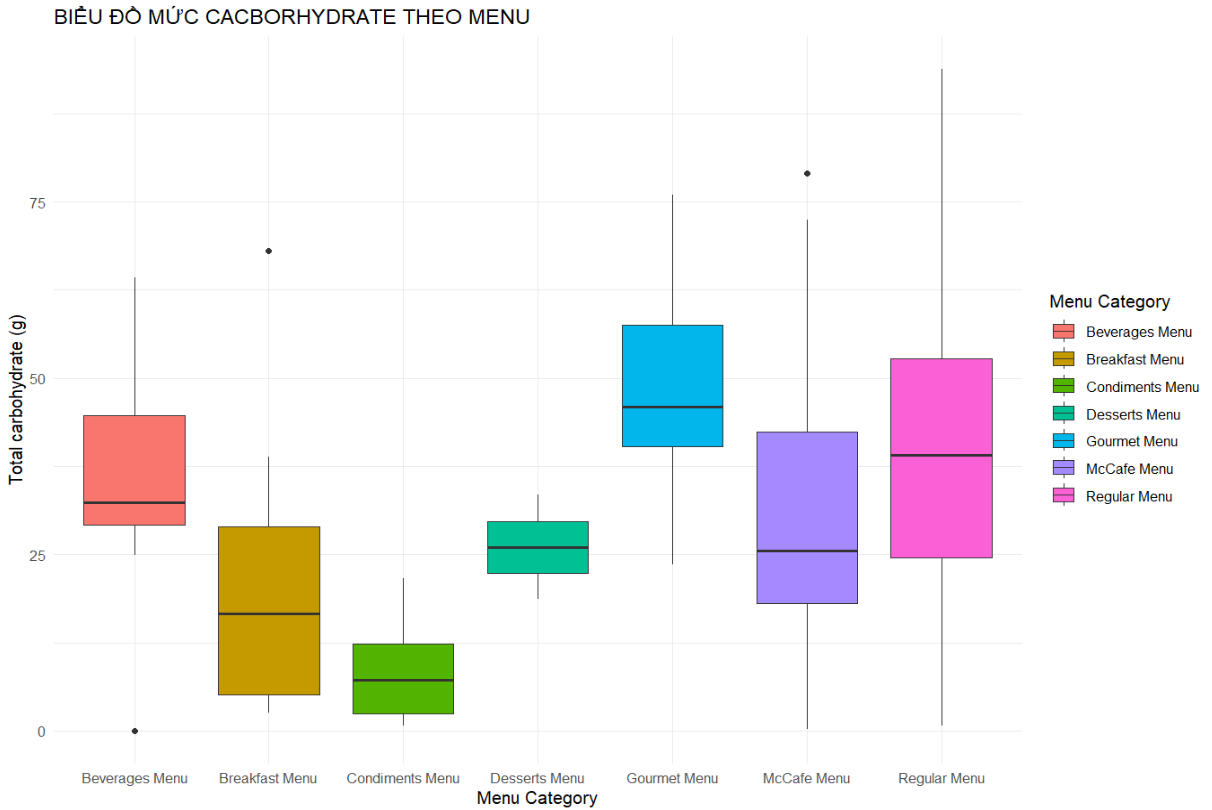
**3.2.4.7. Biểu đồ boxplot của Total carbohydrate**

ggplot(data = data, mapping = aes(y = `Total carbohydrate (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC CARBOHYDRATE THEO MENU")



*Hình 3.27: Biểu đồ boxplot của total carbohydrate*

* Biểu đồ cung cấp cái nhìn trực quan về lượng carbohydrate giữa các loại menu, hữu ích cho việc lập kế hoạch dinh dưỡng hoặc phân tích dinh dưỡng.
* Mỗi thể loại menu có một box plot riêng biệt, cho thấy giá trị trung vị, các phần tư và các điểm ngoại lệ tiềm năng.

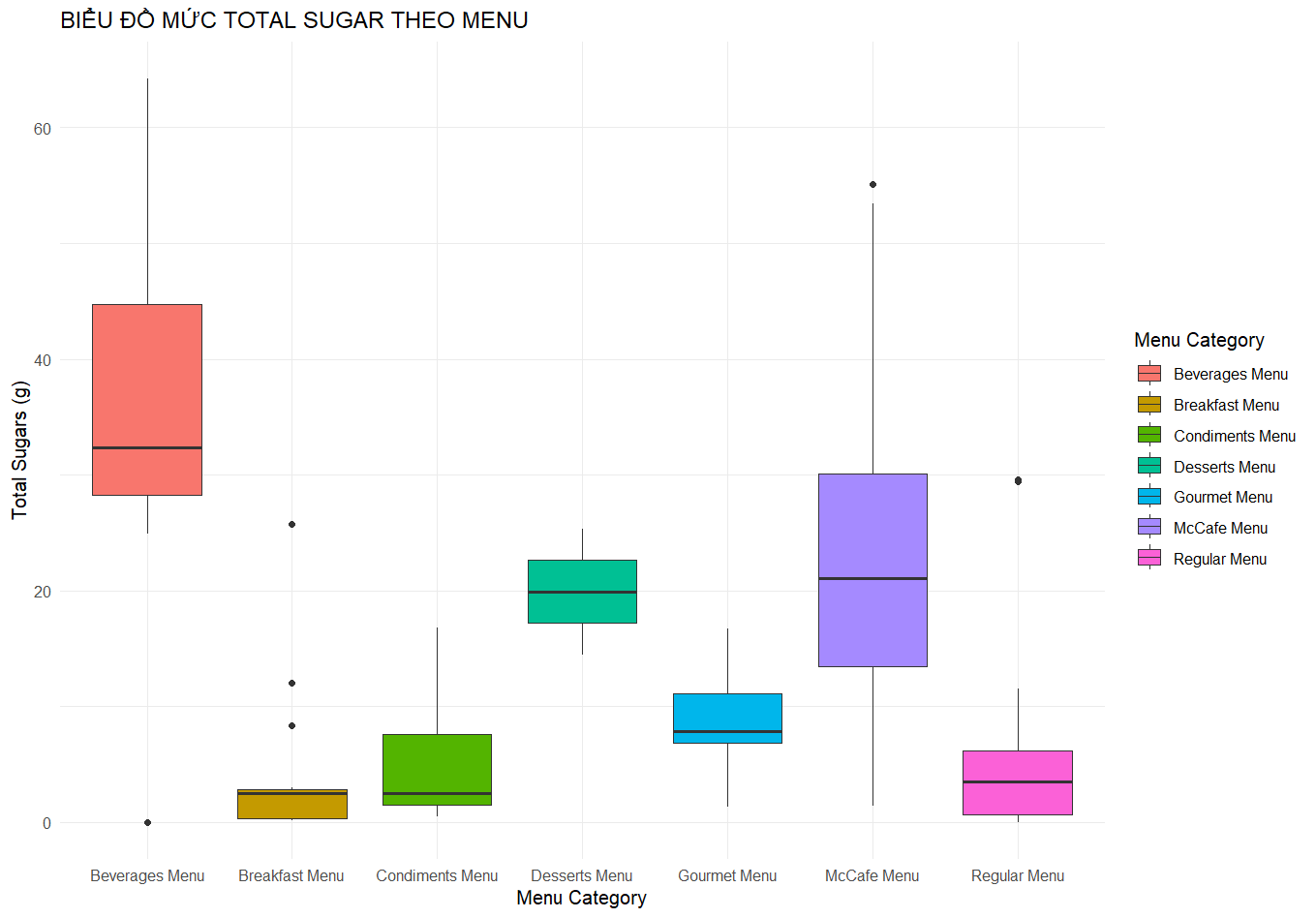
**3.2.4.8. Biểu đồ boxplot của Total sugars**

ggplot(data = data, mapping = aes(y = `Total Sugars (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC TOTAL SUGAR THEO MENU")



*Hình 3.28: Biểu đồ boxplot của total sugars*

* Biểu đồ hộp thể hiện mức độ đường tổng cộng theo từng loại menu tại một chuỗi nhà hàng fast-food.
* Cung cấp một so sánh trực quan về mức độ đường giữa các loại menu khác nhau, có thể chỉ ra thông tin dinh dưỡng quan trọng cho việc xem xét chế độ ăn.

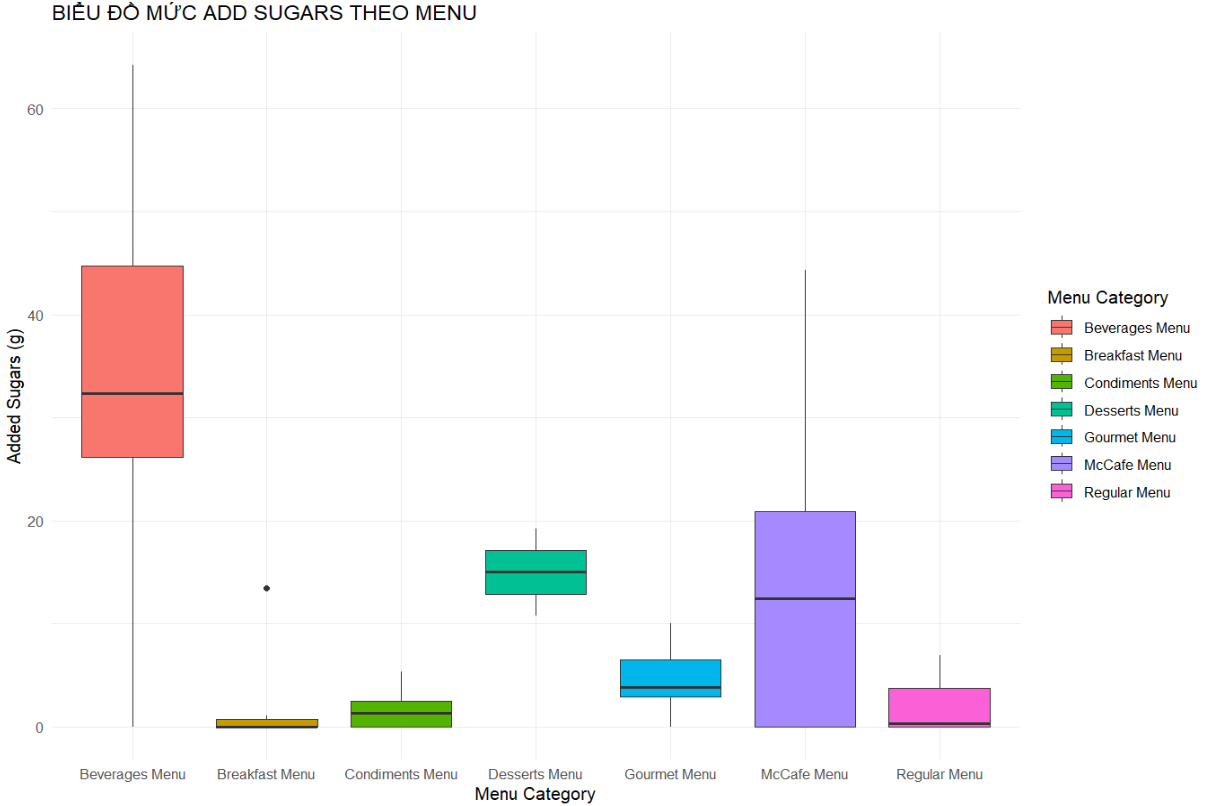
**3.2.4.9. Biểu đồ boxplot của Added sugars**

ggplot(data = data, mapping = aes(y = `Added Sugars (g)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC ADDED SUGARS THEO MENU")



*Hình 3.29: Biểu đồ boxplot của added sugars*

* Biểu đồ hộp này thể hiện mức đường thêm vào (added sugars) theo các danh mục menu khác nhau.
* Beverages Menu và McCafe Menu là hai danh mục có mức đường thêm vào cao nhất và có sự biến động lớn nhất.
* Các danh mục như Breakfast Menu, Condiments Menu và Regular Menu có mức đường thêm vào thấp hơn và ít biến động hơn.
* Mức đường thêm vào trong các danh mục khác (Desserts Menu và Gourmet Menu) nằm ở mức trung bình, với sự biến động nhất định nhưng không quá lớn.

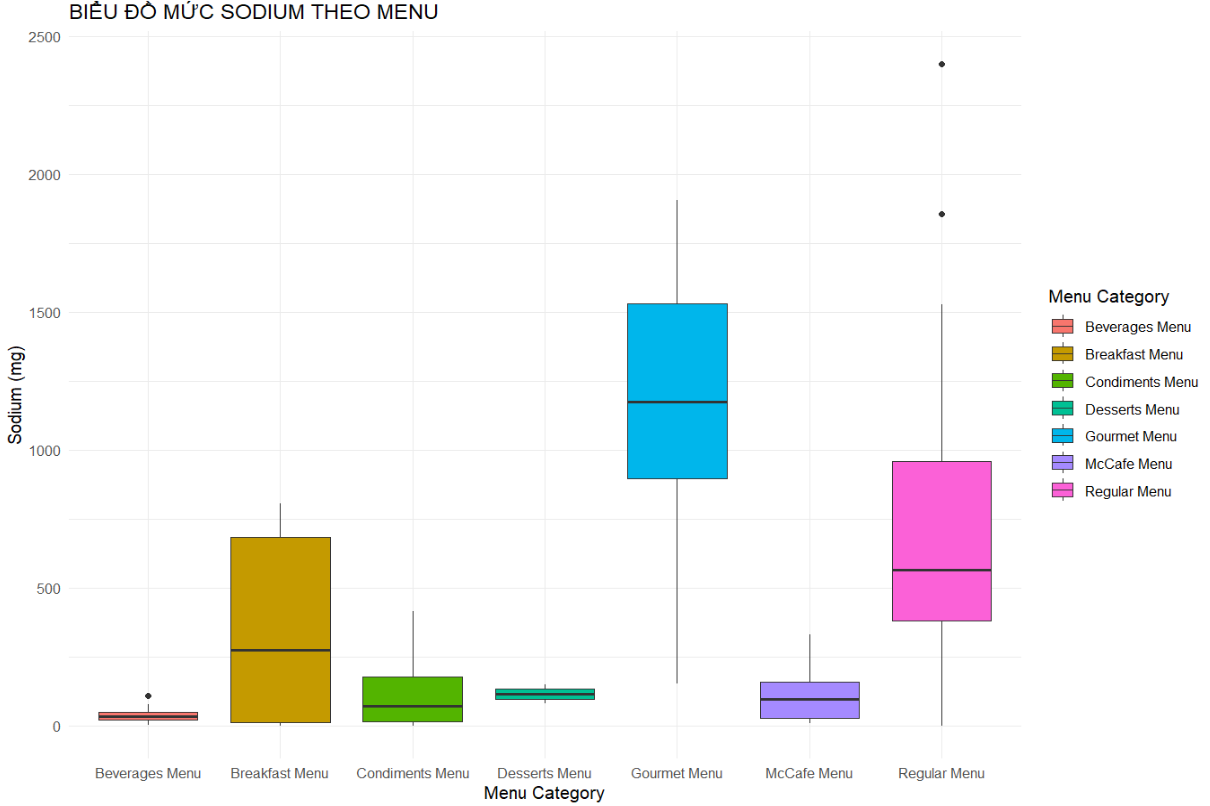
**3.2.4.10. Biểu đồ boxplot của Sodium:**

ggplot(data = data, mapping = aes(y = `Sodium (mg)`, x = `Menu Category`, fill = `Menu Category` ))+

geom\_boxplot()+

theme\_minimal()+

labs(title = "BIỂU ĐỒ MỨC SODIUM THEO MENU")



*Hình 3.30: .Biểu đồ boxplot của sodium*

* Biểu đồ hộp (boxplot) hiển thị mức độ sodium (natri) trong các danh mục thực đơn khác nhau.
* Gourmet Menu và Regular Menu có mức độ sodium cao nhất và phạm vi rộng nhất.
* Beverages Menu, Condiments Menu, và Desserts Menu có mức độ sodium thấp nhất.
* Breakfast Menu có mức độ sodium trung bình với một số giá trị ngoại lệ cao.
* McCafe Menu có mức sodium khá thấp nhưng vẫn có một số giá trị ngoại lệ cao hơn các danh mục khác trừ Regular và Gourmet.
* Biểu đồ cho thấy sự khác biệt đáng kể về mức độ sodium giữa các danh mục thực đơn, từ rất thấp trong đồ uống và tráng miệng đến rất cao trong các món ăn chính và thực đơn cao cấp.

## **3.3. PHÂN TÍCH HỒI QUY TUYẾN TÍNH**

Phân tích hồi quy tuyến tính (linear regression analysis) là một trong những phương pháp phân tích số liệu thông dụng nhất trong thống kê toán học. Có người từng viết “Cho con người 3 vũ khí – hệ số tương quan, hồi quy tuyến tính và một cây bút, con người sẽ sử dụng cả ba”. Phần này sẽ sử dụng R để phân tích hồi quy tuyến tính và vẽ biểu đồ hồi quy tuyến tính.

### **3.3.1. Hệ số tương quan**

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, như giữa "Energy (kCal)"(x) và "Protein (g)"(y). Hệ số tương quan bằng 0 ( hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm (r <0) có nghĩa là khi x tăng cao thì y giảm (và ngược lại, khi x giảm thì y tăng); nếu giá trị hệ số tương quan là dương (r >0) có nghĩa là khi x tăng cao thì y cũng tăng, và khi x giảm cao thì y cũng giảm theo.

Ta có đoạn code sau:

cor(data["Energy (kCal)"],data["Protein (g)"])

cor(data["Energy (kCal)"],data["Total fat (g)"])

cor(data["Energy (kCal)"],data["Sat Fat (g)"])

cor(data["Energy (kCal)"],data["Total carbohydrate (g)"])

cor(data["Energy (kCal)"],data["Sodium (mg)"])

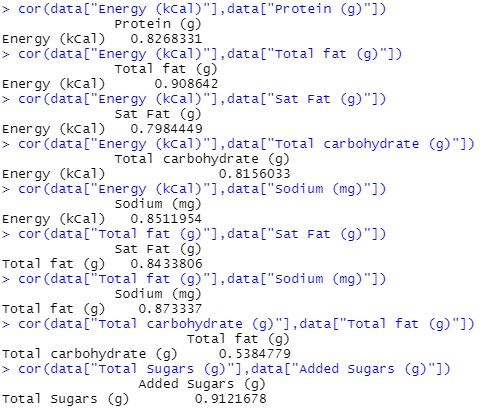
cor(data["Total fat (g)"],data["Sat Fat (g)"])

cor(data["Total fat (g)"],data["Sodium (mg)"])

cor(data["Total carbohydrate (g)"],data["Total fat (g)"])

cor(data["Total Sugars (g)"],data["Added Sugars (g)"])

Ta nhận được bảng sau khi chạy đoạn code này là:



*Bảng 3.1: Kết quả hệ số tương quan giữa các biến*

Dưới đây là một số nhận xét chi tiết về các hệ số tương quan này:

* Energy (kcal)
* Tương quan cao với Total fat (g): 0.908642
* Tương quan tương đối cao với Protein (g): 0.826831
* Tương quan cao với Total carbohydrate (g): 0.815063
* Tương quan với Sodium (mg): 0.851954
* Tương quan với Sat Fat (g): 0.798449
* Total fat (g)
* Tương quan cao với Energy (kcal): 0.908642
* Tương quan rất cao với Sat Fat (g): 0.843806
* Tương quan tương đối cao với Sodium (mg): 0.873397
* Tương quan vừa phải với Total carbohydrate (g): 0.538477
* Tương quan cao với Total Sugars (g): 0.873337
* Total carbohydrate (g)
* Tương quan cao với Energy (kcal): 0.815063
* Tương quan thấp nhất với Total fat (g): 0.538477
* Tương quan thấp với Sodium (mg): 0.614727
* Tương quan rất cao với Total Sugars (g): 0.912678
* Total Sugars (g)
* Tương quan rất cao với Total carbohydrate (g): 0.912678
* Tương quan cao với Total fat (g): 0.873337
* Sat Fat (g)
* Tương quan cao với Total fat (g): 0.843806
* Tương quan tương đối cao với Energy (kcal): 0.798449
* Sodium (mg)
* Tương quan tương đối cao với Energy (kcal): 0.851954
* Tương quan khá cao với Total fat (g): 0.873397
* Từ bảng này, có thể thấy rằng:
* Energy (kcal) có mối tương quan mạnh mẽ với nhiều thành phần dinh dưỡng, đặc biệt là Total fat (g).
* Total Sugars (g) và Total carbohydrate (g) có mối tương quan rất cao, cho thấy chúng có thể liên quan mật thiết với nhau.
* Sat Fat (g) và Total fat (g) cũng có mối tương quan cao, điều này hợp lý vì chất béo bão hòa là một phần của tổng lượng chất béo.

### **3.3.2. Biểu đồ hồi quy**

**3.3.2.1. Biểu đồ hồi quy tuyến tính giữa Energy và Protein**

Tại đây chúng ta sử dụng thư viện ggplot2 để tạo biểu đồ phân tán với đường hồi quy tuyến tính:

ggplot(data, aes(x = `Energy (kCal)`, y = `Protein (g)`)) +

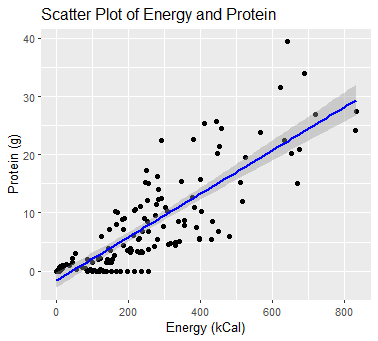
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Energy and Protein",

x = "Energy (kCal)",

y = "Protein (g)")



*Hình 3.31: Biểu đồ hồi quy tuyến tính giữa biến Energy và Protein*

Dưới đây là một số nhận xét chi tiết về biểu đồ này:

* Mối quan hệ giữa Energy và Protein:
* Biểu đồ cho thấy có một xu hướng tăng rõ ràng giữa Energy (kcal) và Protein (g). Khi năng lượng tăng, lượng protein cũng tăng theo. Điều này thể hiện qua đường hồi quy màu xanh dương đi qua các điểm dữ liệu.
* Đường hồi quy có độ dốc dương, cho thấy mối quan hệ tuyến tính dương giữa hai biến này.
* Sự phân tán của dữ liệu:
* Các điểm dữ liệu phân tán khá đều xung quanh đường hồi quy, mặc dù có một số điểm nằm xa đường hồi quy, đặc biệt ở phần năng lượng cao (trên 600 kcal). Điều này có thể cho thấy một số ngoại lệ hoặc biến động trong dữ liệu.
* Mức độ phù hợp của mô hình tuyến tính:
* Độ rộng của dải xám xung quanh đường hồi quy biểu thị khoảng tin cậy của mô hình. Dải này tương đối hẹp, cho thấy mô hình tuyến tính khá phù hợp với dữ liệu và mối tương quan giữa Energy và Protein là khá mạnh.
* Khả năng dự đoán:
* Với sự hiện diện của một số điểm dữ liệu nằm xa khỏi đường hồi quy, có thể thấy rằng mặc dù mô hình tuyến tính là hợp lý cho phần lớn dữ liệu, vẫn có một số trường hợp ngoại lệ cần được xem xét khi dự đoán protein từ năng lượng.
* Hàm aes sẽ thiết lập ánh xạ thẩm mỹ, chỉ định các ánh xạ các biến dữ liệu tới các thuộc tính trực quan. Ở đây `Energy (kCal)` được ánh xạ tới trục x và `Protein (g)` được ánh xạ tới trục y.
* geom\_point() sẽ thêm một lớp điểm vào biểu đồ, tạo thành biểu đồ phân tán. Mỗi điểm biểu thị một quan sát từ dữ liệu có tọa độ x và y được xác định bởi các giá trị `Energy (kCal)` và `Protein (g)`.
* geom\_smooth() sẽ thêm một đường vào biểu đồ, ở đây là đường hồi quy tuyến tính. Với se = TRUE cho biết vùng bóng mờ biểu thị khoảng tin cậy xung quanh đường hồi quy sẽ được hiển thị, nếu đặt thành FALSE khoảng tin cậy sẽ không hiển thị.

**3.3.2.2. Biểu đồ hồi quy tuyến tĩnh giữa Energy và Total fat**

Tương tự như trên ta lần lượt vẽ được các biểu đồ:

ggplot(data, aes(x = `Energy (kCal)`, y = `Total fat (g)`)) +

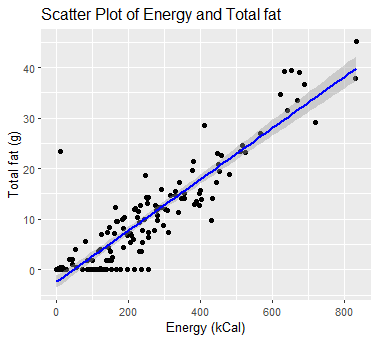
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Energy and Total fat",

x = "Energy (kCal)",

y = "Total fat (g)")



*Hình 3.32: Biểu đồ hồi quy tuyến tính giữa Energy và Total fat*

Biểu đồ này thể hiện mối quan hệ tuyến tính dương giữa Energy (kcal) và Total fat (g). Khi năng lượng tăng, tổng chất béo cũng tăng theo. Đường hồi quy màu xanh dương với dải xám hẹp cho thấy mối tương quan mạnh mẽ và mức độ tin cậy cao. Các điểm dữ liệu phân bố khá đều xung quanh đường hồi quy, ngoại trừ một số ngoại lệ.

**3.3.2.3. Biểu đồ hồi quy tuyến tính giữa Energy và Sat Fat**

ggplot(data, aes(x = `Energy (kCal)`, y = `Sat Fat (g)`)) +

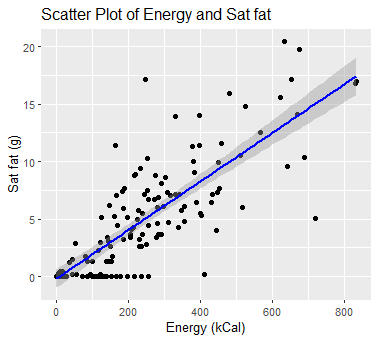
geom\_point() + # Scatter plot

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Energy and Sat fat",

x = "Energy (kCal)",

y = "Sat fat (g)")



*Hình 3.33:Biểu đồ hồi quy tuyến tính giữa Energy và Sat fat*

Biểu đồ scatter plot trên cho thấy mối quan hệ tuyến tính dương giữa năng lượng (kCal) và chất béo bão hòa (Sat fat). Điều này có nghĩa là khi năng lượng trong các món ăn tăng lên, lượng chất béo bão hòa cũng có xu hướng tăng theo. Đường xu hướng màu xanh cùng với dải màu xám thể hiện độ tin cậy của mô hình tuyến tính, cho thấy mối tương quan khá mạnh giữa hai biến này.

**3.3.2.4. Biểu đồ hồi quy tuyến tính giữa Energy và Total carbohydrate**

ggplot(data, aes(x = `Energy (kCal)`, y = `Total carbohydrate (g)`)) +

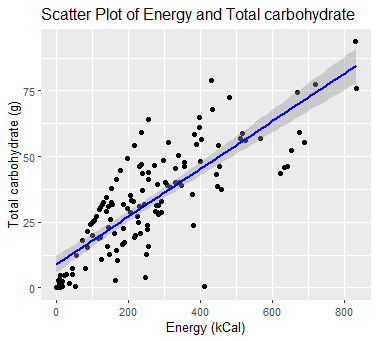
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Energy and Total carbohydrate",

x = "Energy (kCal)",

y = "Total carbohydrate (g)")



*Hình 3.34: Biểu đồ hồi quy tuyến tính giữa Energy và Total carbohydrate*

Biểu đồ scatter plot trên thể hiện mối quan hệ tuyến tính dương giữa năng lượng (kCal) và tổng lượng carbohydrate (g). Điều này cho thấy khi năng lượng trong các món ăn tăng, lượng carbohydrate cũng có xu hướng tăng theo. Đường xu hướng màu xanh cùng với dải màu xám thể hiện độ tin cậy của mô hình tuyến tính, cho thấy mối tương quan khá mạnh giữa hai biến này, với nhiều điểm dữ liệu gần đường xu hướng.

**3.3.2.5. Biểu đồ hồi quy tuyến tính giữa Energy và Sodium**

ggplot(data, aes(x = `Energy (kCal)`, y = `Sodium (mg)`)) +

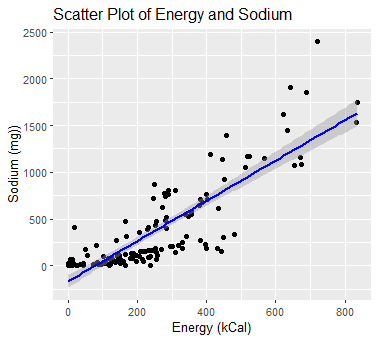
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Energy and Sodium",

x = "Energy (kCal)",

y = "Sodium (mg))")



*Hình 3.35: Biểu đồ hồi quy tuyến tính giữa Energy và Sodium*

Biểu đồ scatter plot này minh họa mối quan hệ tích cực giữa năng lượng (kCal) và sodim (mg) trong các món ăn. Các điểm dữ liệu phân bố cho thấy khi năng lượng tăng, lượng sodium cũng tăng lên. Đường xu hướng màu xanh biểu thị mối quan hệ này, cho thấy một xu hướng tăng rõ ràng giữa hai biến.

**3.3.2.6. Biểu đồ hồi quy tuyến tính giữa Protein và Total fat**

ggplot(data, aes(x = `Protein (g)`, y = `Total fat (g)`)) +

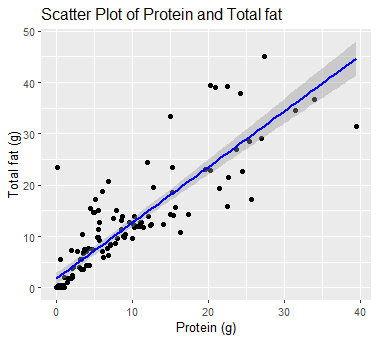
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Protein and Total fat",

x = "Protein (g)",

y = "Total fat (g)")



*Hình 3.36: Biểu đồ hồi quy tuyến tính giữa Protein và Total fat*

Biểu đồ scatter plot này cho thấy mối quan hệ tích cực giữa hàm lượng protein và tổng lượng chất béo trong các món ăn. Cụ thể:

* Xu hướng tích cực: Khi hàm lượng protein tăng, tổng lượng chất béo cũng tăng theo, thể hiện qua đường hồi quy màu xanh dương với dải tin cậy màu xám.
* Phân tán dữ liệu: Các điểm dữ liệu tập trung gần đường hồi quy, chỉ ra mối tương quan chặt chẽ giữa protein và chất béo.
* Điểm ngoại lệ: Có một vài điểm nằm xa đường hồi quy, cho thấy một số món ăn có hàm lượng protein và chất béo không theo xu hướng chung.

**3.3.2.7. Biểu đồ hồi quy tuyến tính giữa Protein và Sodium**

ggplot(data, aes(x = `Protein (g)`, y = `Sodium (mg)`)) +

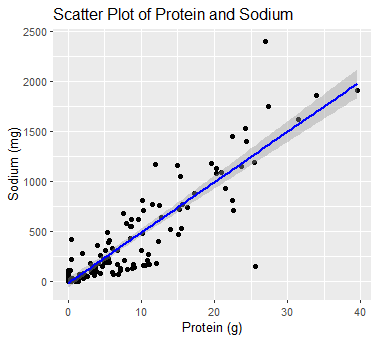
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Protein and Sodium",

x = "Protein (g)",

y = "Sodium (mg)")



*Hình 3.37: Biểu đồ hồi quy tuyến tính giữa Protein và Sodium*

Biểu đồ này cho thấy:

* Xu hướng tích cực: Protein tăng thì sodium cũng tăng.
* Tương quan chặt: Các điểm dữ liệu tập trung quanh đường hồi quy.
* Điểm ngoại lệ: Một số món ăn không theo xu hướng chung.

**3.3.2.8. Biểu đồ hồi quy tuyến tính giữa Total fat và Sat fat**

ggplot(data, aes(x = `Total fat (g)`, y = `Sat Fat (g)`)) +

geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Total fat and Sat fat",

x = "Total fat (g)",

y = "Sat fat (g)")



*Hình 3.38: Biểu đồ hồi quy tuyến tính giữa Total fat và Sat fat*

Biểu đồ scatter plot này thể hiện mối quan hệ giữa tổng lượng chất béo (Total fat) và chất béo bão hòa (Sat fat) trong các món ăn.

* Xu hướng tích cực: Khi tổng lượng chất béo tăng, chất béo bão hòa cũng tăng.
* Tương quan chặt: Các điểm dữ liệu tập trung quanh đường hồi quy, cho thấy mối tương quan chặt chẽ.
* Điểm ngoại lệ: Một số điểm nằm xa đường hồi quy, cho thấy một số món ăn không theo xu hướng chung.

**3.3.2.9. Biểu đồ hồi quy tuyến tính giữa Total fat và Sodium**

ggplot(data, aes(x = `Total fat (g)`, y = `Sodium (mg)`)) +

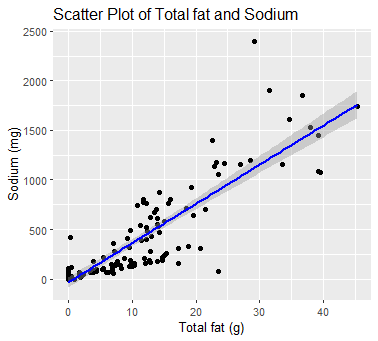
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Total fat and Sodium",

x = "Total fat (g)",

y = "Sodium (mg)")



*Hình 3.39: Biểu đồ hồi quy tuyến tính giữa Total fat và Sodium*

Biểu đồ tán xạ cho thấy mối quan hệ tuyến tính dương giữa tổng lượng chất béo và lượng sodium, tức là khi lượng chất béo tăng, lượng sodium cũng tăng. Mặc dù có một số điểm ngoại lệ, phần lớn dữ liệu nằm gần đường hồi quy, cho thấy mối quan hệ khá mạnh mẽ giữa hai biến số này.

**3.3.2.10. Biểu đồ hồi quy tuyến tính giữa Total carbohydrate và Total fat**

ggplot(data, aes(x = `Total carbohydrate (g)`, y = `Total fat (g)`)) +

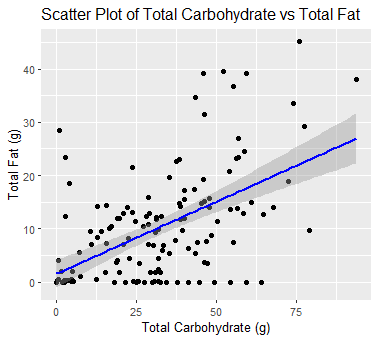
geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Total Carbohydrate vs Total Fat",

x = "Total Carbohydrate (g)",

y = "Total Fat (g)")



*Hình 3.40: Biểu đồ hồi quy tuyến tính giữa Total carbohydrate và Total fat*

Biểu đồ tán xạ này minh họa mối quan hệ dương giữa tổng lượng carbohydrate (g) và tổng lượng chất béo (g). Đường hồi quy màu xanh cho thấy khi lượng carbohydrate tăng, lượng chất béo cũng tăng. Mặc dù có sự phân tán dữ liệu, xu hướng tổng quát vẫn là tích cực.

**3.3.2.11. Biểu đồ hồi quy tuyến tính giữa Total Sugars và Added Sugars**

ggplot(data, aes(x = `Total Sugars (g)`, y = `Added Sugars (g)`)) +

geom\_point() +

geom\_smooth(method = "lm", se = TRUE, color = "blue") +

labs(title = "Scatter Plot of Total Sugars vs Added Sugars",

x = "Total Sugars (g)",

y = "Added Sugars (g)")



*Hình 3.41: Biểu đồ hồi quy tuyến tính giữa Total Sugars và Added Sugars*

Biểu đồ tán xạ này minh họa mối quan hệ tích cực chặt chẽ giữa tổng lượng đường (g) và lượng đường thêm vào (g). Đường hồi quy màu xanh cho thấy khi tổng lượng đường tăng, lượng đường thêm vào cũng tăng tương ứng. Mối tương quan giữa hai biến số này rất rõ ràng và mạnh mẽ.

## **3.4. PHÂN TÍCH THÀNH PHẦN CHÍNH PCA**

Phân tích thành phần chính (PCA) cho phép chúng ta tóm tắt và hiển thị thông tin trong một bộ dữ liệu chứa các cá nhân/quan sát được mô tả bởi nhiều biến số định lượng tương quan với nhau. Mỗi biến số có thể được coi là một chiều khác nhau. Nếu bạn có nhiều hơn 3 biến số trong bộ dữ liệu của mình, việc hiển thị một không gian siêu chiều nhiều chiều có thể rất khó khăn.

Phân tích thành phần chính được sử dụng để trích xuất thông tin quan trọng từ một bảng dữ liệu đa biến và biểu diễn thông tin này dưới dạng một tập hợp ít biến mới gọi là các thành phần chính. Các biến mới này tương ứng với một tổ hợp tuyến tính của các biến ban đầu. Số lượng thành phần chính ít hơn hoặc bằng số lượng biến ban đầu.

Thông tin trong một bộ dữ liệu cụ thể tương ứng với biến thể tổng cùng chứa. Mục tiêu của PCA là xác định các hướng (hoặc thành phần chính) theo đó biến thể trong dữ liệu là tối đa.

Nói cách khác, PCA giảm chiều của một bộ dữ liệu đa biến xuống còn hai hoặc ba thành phần chính, có thể được hiển thị đồ họa, với sự mất mát thông tin tối thiểu.

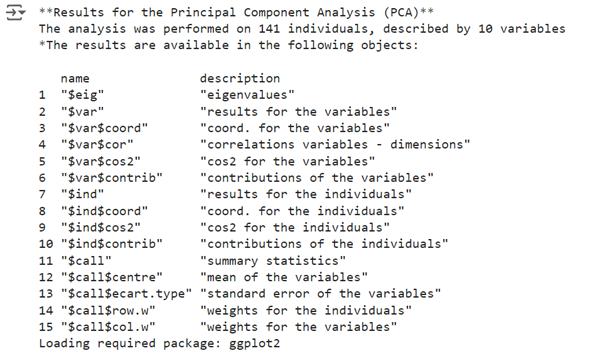
### **3.4.1. Chuẩn hóa dữ liệu**

library("FactoMineR")

PCA\_data <- data[,4:13]

res.pca <- PCA(PCA\_data, graph = FALSE)

print(res.pca)



*Bảng 3.2: Kết quả phân tích thành phần chính (PCA)*

Đối tượng được tạo bằng hàm PCA () chứa nhiều thông tin được tìm thấy trong nhiều danh sách và ma trận khác nhau. Những giá trị này được mô tả trong phần tiếp theo.

### **3.4.2. Trực quan hóa và giải thích**

**3.4.2.1. Giá trị riêng / Phương sai**

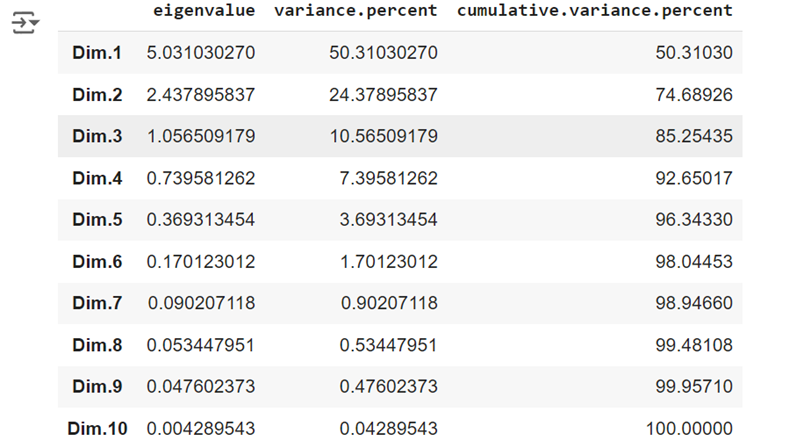
Giá trị riêng đo lường mức độ biến thiên được giữ lại bởi mỗi thành phần chính. Giá trị riêng lớn đối với các PC đầu tiên và nhỏ đối với các PC tiếp theo. Nghĩa là, các PC đầu tiên tương ứng với các hướng có mức độ biến đổi tối đa trong tập dữ liệu.

Chúng ta kiểm tra các giá trị riêng để xác định số lượng thành phần chính cần xem xét. Các giá trị riêng và tỷ lệ phương sai (tức là thông tin) được giữ lại bởi các thành phần chính (PC) có thể được trích xuất bằng cách sử dụng hàm get\_eigenvalue() [ gói Factoextra ].

library("factoextra")

eig.val <- get\_eigenvalue(res.pca)

eig.val



*Bảng 3.3: Giá trị riêng*

* Tổng của tất cả các giá trị riêng cho tổng phương sai là 10.Tỷ lệ biến thiên được giải thích bởi mỗi giá trị riêng được cho trong cột thứ hai.

***Ví dụ***: 50,3 chia cho 10 bằng 5.03, hoặc khoảng 50,3% biến thể được giải thích bằng giá trị riêng đầu tiên này.

* Tỷ lệ phần trăm tích lũy được giải thích thu được bằng cách cộng các tỷ lệ biến thiên liên tiếp được giải thích để thu được tổng số đang chạy.

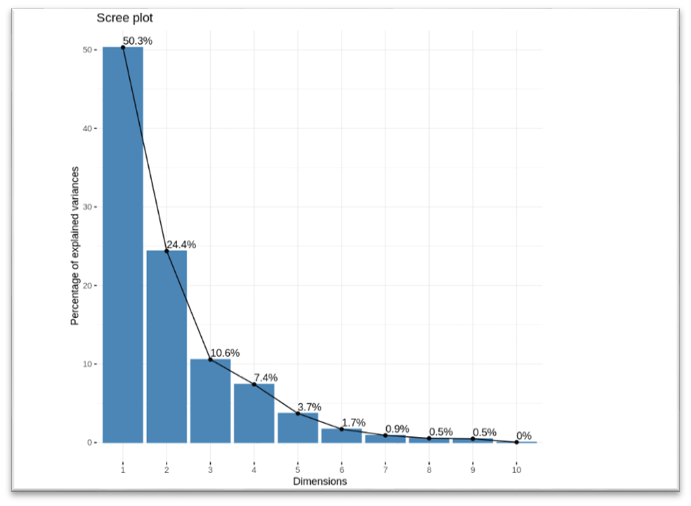
***Ví dụ***: 50,31% cộng với 24,37% bằng 74,68%, v.v. Do đó, khoảng 74,68% biến thiên được giải thích bởi hai giá trị riêng đầu tiên cộng lại.

* Trong phân tích này, ba thành phần chính đầu tiên giải thích 84% sự khác biệt. Đây là một tỷ lệ lớn có thể chấp nhận được.

Một phương pháp khác để xác định số lượng thành phần chính là xem “Biểu đồ Scree”, là biểu đồ các giá trị riêng được sắp xếp từ lớn nhất đến nhỏ nhất. Số lượng thành phần được xác định tại điểm, ngoài điểm đó các giá trị riêng còn lại đều tương đối nhỏ và có kích thước tương đương.

Scree plot có thể được tạo bằng cách sử dụng hàm fviz\_eig() hoặc fviz\_screeplot()[ gói Factoextra ].

fviz\_eig(res.pca, addlabels = TRUE, ylim = c(0, 50))



*Hình 3.42: Biểu đồ scree plot thể hiện giá trị riêng*

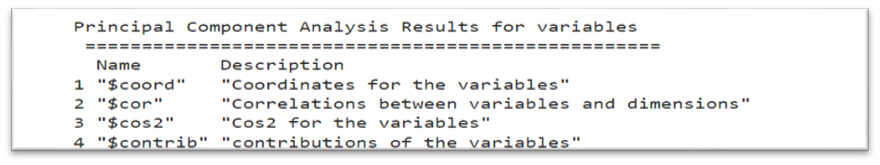
Từ biểu đồ trên, chúng ta có thể muốn dừng lại ở thành phần chính thứ tư. 92,7% thông tin (phương sai) có trong dữ liệu được bốn thành phần chính đầu tiên giữ lại.

### **3.4.3. Đồ thị của các biến**

Một phương pháp đơn giản để trích xuất kết quả cho các biến từ đầu ra PCA là sử dụng hàm get\_pca\_var()[ gói Factoextra ]. Hàm này cung cấp danh sách các ma trận chứa tất cả các kết quả cho các biến hoạt động (tọa độ, mối tương quan giữa các biến và trục, cosin bình phương và đóng góp)

var <- get\_pca\_var(res.pca)

var



*Bảng 3.4: Kết quả phân tích thành phần chính cho các biến*

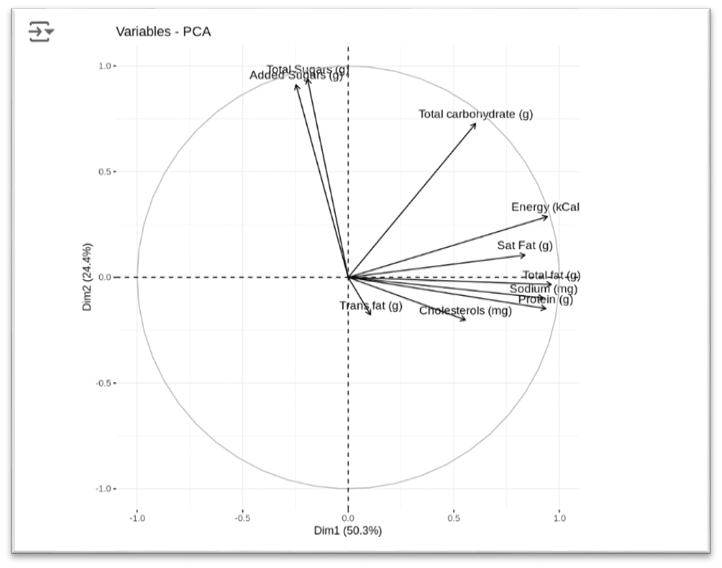
Các thành phần của get\_pca\_var()có thể được sử dụng trong biểu đồ biến như sau:

* var$coord: tọa độ của các biến để tạo biểu đồ phân tán
* var$cos2: thể hiện chất lượng biểu diễn của các biến trên bản đồ nhân tố. Nó được tính theo tọa độ bình phương: var.cos2 = var.coord \* var.coord.
* var$contrib: chứa sự đóng góp (tính bằng phần trăm) của các biến đối với các thành phần chính. Sự đóng góp của một biến (var) cho một thành phần chính nhất định là (tính bằng phần trăm): (var.cos2 \* 100) / (tổng cos2 của thành phần).

**3.4.3.1. Vòng tròn tương quan**

Sự tương quan giữa một biến và một thành phần chính (PC) được sử dụng như là tọa độ của biến trên PC. Cách biểu diễn của các biến khác biệt so với biểu đồ của các quan sát: Các quan sát được biểu diễn bằng các chiếu, nhưng các biến lại được biểu diễn bằng các tương quan.

fviz\_pca\_var(res.pca, col.var = "black")



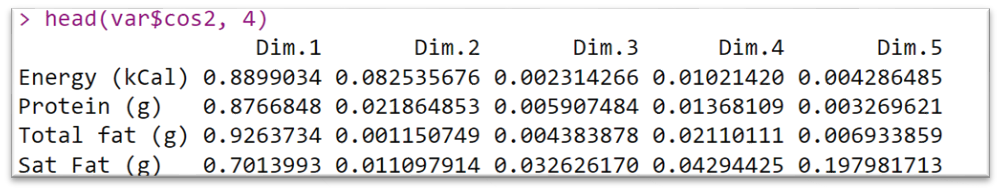
*Hình 3.43: Vòng tròn tương quan của các biến*

* Đồ thị trên còn được gọi là đồ thị tương quan biến đổi. Nó cho thấy mối quan hệ giữa tất cả các biến. Nó có thể được hiểu như sau:
* Các biến tương quan dương được nhóm lại với nhau.
* Các biến tương quan âm được định vị ở các phía đối diện của gốc đồ thị (góc phần tư đối lập).
* Khoảng cách giữa các biến và gốc đo lường chất lượng của các biến trên bản đồ nhân tố. Các biến nằm xa nguồn gốc được thể hiện tốt trên bản đồ nhân tố.

**3.4.3.2. Chất lượng đại diện**

Chất lượng biểu diễn của các biến trên bản đồ yếu tố được gọi là cos2 (cosin vuông, tọa độ bình phương) . Bạn có thể truy cập vào cos2 như sau:

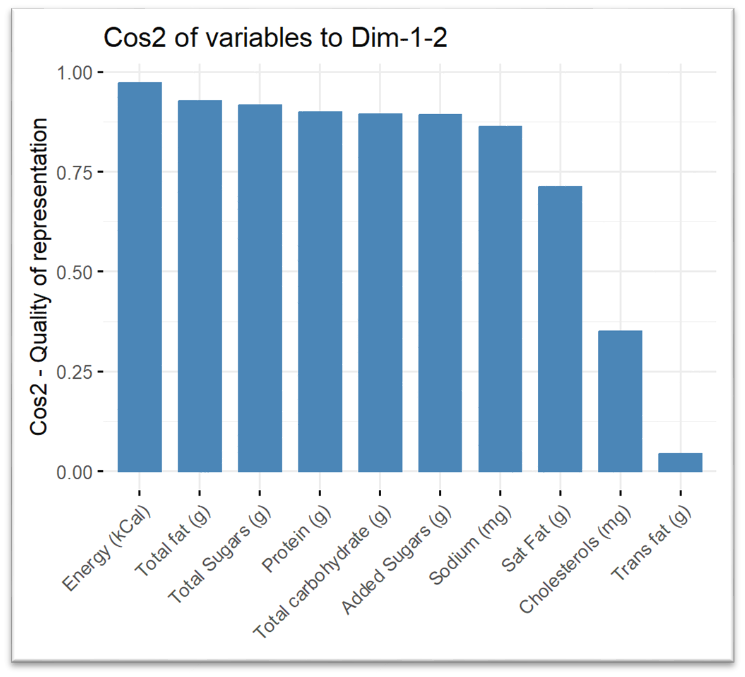
head(var$cos2, 4)



*Bảng 3.5: 4 dòng của var$cos2*

# Total cos2 of variables on Dim.1 and Dim.2

fviz\_cos2(res.pca, choice = "var", axes = 1:2)



*Hình 3.44: Biểu đồ thể hiện chất lượng thể hiện của 2 thành phần chính*

Từ biểu đồ ta thấy được chất lượng biểu diễn lớn của 7 biến đầu tiên bao gồm Energy, total fat, total sugar, protein, total carbohydrate, added sugar, sodium, với mỗi biến đóng góp hơn 75% vào thành phần chính thứ 1 và thứ 2.

Tóm tắt:

* Các giá trị cos2 được sử dụng để ước tính chất lượng của biểu diễn.
* Một biến càng gần với vòng tròn tương quan, biểu diễn của nó trên bản đồ yếu tố càng tốt (và việc giải thích các thành phần này càng quan trọng).
* Các biến được đóng vào trung tâm của biểu đồ ít quan trọng hơn đối với các thành phần đầu tiên.
* Màu theo giá trị cos2: chất lượng trên bản đồ yếu tố

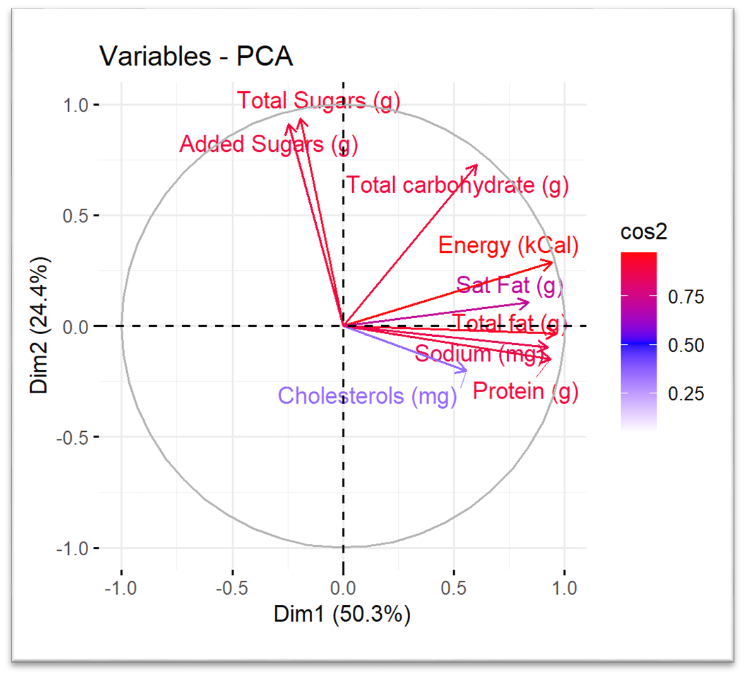
# Color by cos2 values: quality on the factor map

fviz\_pca\_var(res.pca, col.var = "cos2",

gradient.cols = c("white", "blue", "red"),

repel = TRUE # Avoid text overlapping

)



*Hình 3.45 : Biểu đồ vòng tròn tương quan thể hiện màu sắc theo giá trị*

Từ vòng tròn tương quan theo màu sắc ta thấy được các chất lượng biểu diễn ta đánh giá bên trên đều đóng góp rất lớn của giá trị thể hiện qua màu sắc :

* Các biến có giá trị COS2 thấp sẽ được tô màu "trắng”.
* Các biến có giá trị COS2 trung bình sẽ được tô màu "xanh lam".
* Các biến có giá trị COS2 cao sẽ được tô “màu đỏ”.

# **CHƯƠNG 4: TỔNG KẾT**

Việc trực quan hóa các thành phần dinh dưỡng của McDonald's qua nhiều loại biểu đồ khác nhau, bao gồm biểu đồ tán xạ, biểu đồ hộp, biểu đồ thanh, cùng với thống kê mô tả và ma trận tương quan, đã cung cấp cái nhìn toàn diện và sâu sắc về mối quan hệ giữa các yếu tố dinh dưỡng quan trọng.

Các biểu đồ tán xạ minh họa mối quan hệ tuyến tính dương giữa các thành phần dinh dưỡng như tổng lượng chất béo, carbohydrate và đường. Điều này cho thấy khi một yếu tố dinh dưỡng tăng, các yếu tố khác cũng có xu hướng tăng theo. Biểu đồ hộp giúp chúng ta hiểu rõ hơn về sự phân bố của từng thành phần dinh dưỡng, thể hiện rõ ràng các giá trị trung vị, tứ phân vị và các điểm ngoại lệ, từ đó xác định các món ăn có giá trị dinh dưỡng đặc biệt cao hoặc thấp. Biểu đồ thanh cung cấp một cái nhìn trực quan về mức độ phổ biến của các thành phần dinh dưỡng trong các món ăn, giúp dễ dàng so sánh lượng chất béo, carbohydrate và đường giữa các sản phẩm.

Thống kê mô tả đã cung cấp các chỉ số cụ thể như giá trị trung bình, độ lệch chuẩn, và phạm vi của từng thành phần dinh dưỡng. Những con số này cho thấy đặc điểm chung và sự biến thiên của các món ăn tại McDonald's, giúp người tiêu dùng hiểu rõ hơn về giá trị dinh dưỡng của thực phẩm họ tiêu thụ. Ma trận tương quan giữa các thành phần dinh dưỡng cho thấy mức độ liên hệ giữa các biến số, chẳng hạn như mối quan hệ mật thiết giữa chất béo và carbohydrate. Điều này có thể giúp xác định xu hướng dinh dưỡng chung của các món ăn.

Những thông tin này không chỉ hữu ích cho người tiêu dùng trong việc đưa ra lựa chọn thực phẩm thông minh mà còn cung cấp dữ liệu quan trọng cho các nhà nghiên cứu và chuyên gia dinh dưỡng. Họ có thể sử dụng các kết quả này để phân tích xu hướng dinh dưỡng, phát triển các khuyến nghị ăn uống hợp lý, và cải thiện chất lượng dinh dưỡng của các món ăn.

Tóm lại, việc sử dụng đa dạng các phương pháp trực quan hóa và phân tích thống kê đã mang lại cái nhìn toàn diện và chi tiết về thành phần dinh dưỡng của McDonald's. Điều này không chỉ giúp nâng cao nhận thức của người tiêu dùng về giá trị dinh dưỡng mà còn hỗ trợ các nhà nghiên cứu trong việc đưa ra các đề xuất cải thiện chế độ ăn uống, góp phần thúc đẩy lối sống lành mạnh và lựa chọn thực phẩm thông minh.

# **TÀI LIỆU THAM KHẢO**

[1]. Phạm Duy Hùng, Đỗ Thái Giang (2022), *Ngôn ngữ lập trình R*, Nhà xuất bản ĐH Quốc gia Hà Nội.

[2]. Trần Thị Kim Thanh, Trần Chí Lê (2023), *Tài liệu học tập Thống kê Toán học cho ngành Khoa học dữ liệu*, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[3]. Nguyễn Văn Tuấn (2020), *Phân tích dữ liệu với* ***R***, Nhà xuất bản tổng hợp TP. Hồ Chí Minh.

[4]. Nguyễn Văn Tuấn (2019), *Phân tích dữ liệu với* ***R****, hỏi và đáp*, Nhà xuất bản tổng hợp TP. Hồ Chí Minh.

[5]. Trần Chí Lê, Nguyễn Thị Hạnh Lê (2024), *Tài liệu học tập Đồ án 1: Trực quan hóa dữ liệu bằng* ***R***, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[6]. Link tham khảo:

<https://www.kaggle.com/datasets/deepcontractor/mcdonalds-india-menu-nutrition->

facts

[7]. Web: <https://machinelearningcoban.com/2017/06/15/pca/>

[8]. Web: <https://rpubs.com/phucle98/habeco1>

# 

# 

# 