BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

**BÁO CÁO ĐỒ ÁN 1**

**SỬ DỤNG R ĐỂ PHÂN TÍCH VÀ VẼ BIỂU ĐỒ DỮ LIỆU XĂNG DẦU PETROLIMEX**

|  |  |
| --- | --- |
| HỌ VÀ TÊN SINH VIÊN THỰC HIỆN ĐỒ ÁN 1 | MÃ SINH VIÊN |
| LÊ QUANG LINH | 22174600061 |
| NGUYỄN HỮU ĐỨC | 22174600048 |
| TỪ TẤT DUY | 22174600050 |
| TRẦN ANH LÃM | 22174600064 |

GIẢNG VIÊN HƯỚNG DẪN

TRẦN CHÍ LÊ

HÀ NỘI – 2024

**MỤC LỤC**

[**CHƯƠNG 1: GIỚI THIỆU TỔNG QUAN VỀ THƯ VIỆN GGPLOT VÀ PHÂN TÍCH THỐNG KÊ** 3](#_Toc167671994)

[**1. THĂM DÒ DỮ LIỆU BẰNG BIỂU ĐỒ** 3](#_Toc167671995)

[1.1 BIỂU ĐỒ VỚI PACKAGE GGPLOT2 3](#_Toc167671996)

[1.2. PHÂN TÍCH THỐNG KÊ 23](#_Toc167671997)

[**CHƯƠNG 2: PHÂN TÍCH THĂM DÒ DỮ LIỆU** 27](#_Toc167671998)

[**1. PHƯƠNG PHÁP PHÂN TÍCH THĂM DÒ DỮ LIỆU** 27](#_Toc167671999)

[1.1 GIỚI THIỆU 27](#_Toc167672000)

[1.2 THỐNG KÊ MÔ TẢ (DESCRIPTIVE STATISTICS) 27](#_Toc167672001)

[1.3 TRỰC QUAN HÓA DỮ LIỆU (DATA VISUALIZATION) 28](#_Toc167672002)

[1.4 PHÂN TÍCH TƯƠNG QUAN (CORRELATION ANALYSIS) 31](#_Toc167672003)

[1.5 PHÁT HIỆN DỮ LIỆU NGOẠI LỆ (OUTLIER DETECTION) 31](#_Toc167672004)

[1.6 PHÂN TÍCH PHÂN PHỐI (DISTRIBUTION ANALYSIS) 33](#_Toc167672005)

[**2. CÔNG CỤ SỬ DỤNG TRONG EDA** 35](#_Toc167672006)

[2.1 R 35](#_Toc167672007)

[2.2 MICROSOFT EXCEL 35](#_Toc167672008)

[2.3 TRỰC QUAN HÓA BIẾN ĐỘNG GIÁ XĂNG DẦU 35](#_Toc167672009)

[2.4 CHUẨN HÓA DỮ LIỆU 36](#_Toc167672010)

[2.5 TÍNH TOÁN MA TRẬN 36](#_Toc167672011)

[**CHƯƠNG 3: TRỰC QUAN DỮ LIỆU** 37](#_Toc167672012)

[**1. PHÂN TÍCH THĂM DÒ DỮ LIỆU THAY ĐỔI XĂNG DẦU VIỆT NAM TRONG NĂM 2023 VÀ 2024 VỚI BIỂU ĐỒ PHÂN PHỐI CHUẨN** 37](#_Toc167672013)

[1.1 GIỚI THIỆU 37](#_Toc167672014)

[1.2 ĐỌC FILE DỮ LIỆU XĂNG DẦU 38](#_Toc167672015)

[1.3 NHẬN XÉT 46](#_Toc167672016)

[**2. PHÂN TÍCH THÀNH PHẦN CHÍNH -PCA** 46](#_Toc167672017)

[2.1 GIỚI THIỆU 46](#_Toc167672018)

[2.1 KHÁI NIỆM 48](#_Toc167672019)

[2.2 ĐẶC TÍNH PCA 49](#_Toc167672020)

[2.3 MÔ HÌNH PCA 49](#_Toc167672021)

[2.4 THỰC HÀNH PHÂN TÍCH PCA TRONG R VỚI DỮ LIỆU XĂNG DẦU PETROLIMEX: 49](#_Toc167672022)

[**KẾT LUẬN** 64](#_Toc167672023)

[**TÀI LIỆU THAM KHẢO** 64](#_Toc167672024)

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

**CÁC KỸ THUẬT THĂM DÒ VÀ TRỰC QUAN HÓA DỮ LIỆU CƠ BẢN TRONG R**

Tóm tắt kiến thức: Bài này giới thiệu các kỹ thuật phân tích thống kê (ước lượng, kiểm định các tham số của các số liệu mẫu dựa trên quy luật phân phối tương ứng) và kỹ thuật xây dựng biểu đồ cho từng đối tượng dữ liệu (về độ lớn dữ liệu, dữ liệu biểu thị dạng phần trăm, về quy luật phân phối của dữ liệu và mối tương quan giữa các nhóm biến trong dữ liệu), (xem [1,2,14,15]).

### **1. THĂM DÒ DỮ LIỆU BẰNG BIỂU ĐỒ**

### 1.1 BIỂU ĐỒ VỚI PACKAGE GGPLOT2

**

Hình 1.1: Minh họa qui 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 gói package *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).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| mpg #>  #> | # | A tibble: 234 × 11 manufacturer model | | displ | year | cyl trans | drv | cty | hwy fl class | | |
| #> |  | <chr> <chr> | | <dbl> | <int> | <int> <chr> | <chr> | <int> | <int> <chr> <chr> | | |
| #> | 1 | audi a4 | | 1.8 | 1999 | 4 auto(l5) | f | 18 | 29 p compa… | | |
| #> | 2 | audi | a4 | 1.8 | 1999 | 4 manual(m5) f | | 21 | 29 | p | compa… |
| #> | 3 | audi | a4 | 2 | 2008 | 4 manual(m6) f | | 20 | 31 | p | compa… |
| #> | 4 | audi | a4 | 2 | 2008 | 4 auto(av) f | | 21 | 30 | p | compa… |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| #> | 5 | audi | a4 | 2.8 1999 | 6 auto(l5) | f | 16 | 26 p | compa… |
| #>  #> | 6  # | audi  ℹ 228 | a4  more rows | 2.8 1999 | 6 manual(m5) | f | 18 | 26 p | compa… |

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

1. *Cú pháp cơ bản*

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

ggplot(data = my\_data)+ # sừ dụng dữ li̘u "my\_data" geom\_yyy( # thêm một lỚp các hàm-hình biểu đồ

mapping = aes(x = col1, y = col2), # gán dữ li̘u tỚi các trục

color = "red")+ # thêm một số đ̐c điểm khác (nhG màu sắc) labs()+ # thêm tiêu đề, nhãn, bàng số,..

theme() # điều chình cỠ chữ, màu sắc, phông chữ

1. *Gán các biến dũ li͏u cho biểu đồ*

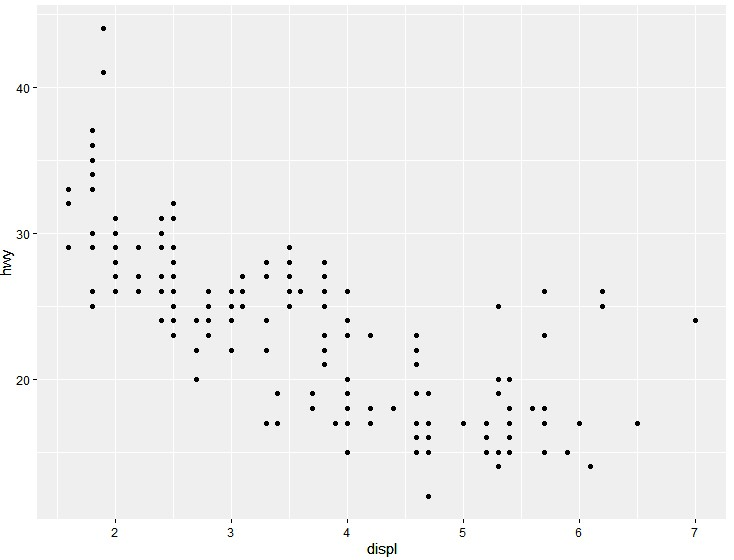
Hầu hết các hàm-hình geom phải được cho biết cái gì được sử dụng để vẽ biểu đồ, vì vậy chúng ta phải cung cấp cách map (gán) các biến số trong dữ liệu tới các thành phần của biểu đồ như là các trục, màu đối tượng, kích thước đối tượng, v.v. Đối với hầu hết các geoms, các thành phần thiết yếu phải được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

***V****í* ***d****ụ* ***1.1.*** Trong lệnh ggplot() dưới đây, dữ liệu được thiết lập là bộ dữ liệu mpg. Trong

đó đối số mapping = aes(), cột displ được gán cho trục x, và cột hwy được gán cho trục y.

ggplot(data = my\_data)+

geom\_poit(mapping = aes(x = displ, y = hwy)) # kết quà hiển th$:

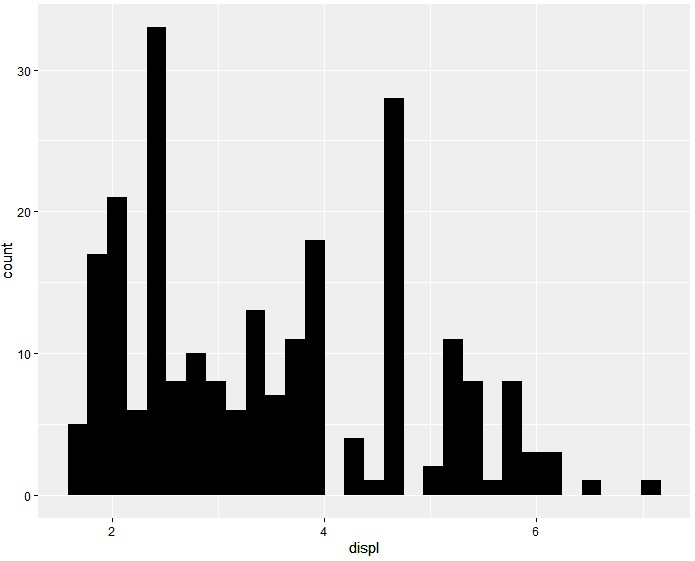


Hình 1.2: Cách gán biến số vào các trục.

***V****í* ***d****ụ* ***1.2.*** Lệnh sau tương tự Ví dụ 1.1, chỉ có một sự khác biệt nhỏ về cách mapping và hàm geom. Hàm geom\_histogram() chỉ yêu cầu gán cột cho trục x, bởi vì trục số lượng y được tạo ra một cách tự động bằng số phần tử của mỗi biến x.

ggplot(data = mpg, mapping = aes(x = displ))+ geom\_histogram()

# kết quà hiển th$:



Hình 1.3: Cách gán biến vào lớp hàm geom\_histogram.

1. *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 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 gһp:

* + shape = Hiển thị một điểm với hàm geom\_point() dưới dạng dấu chấm, ngôi sao, hình tam giác hoһc hình vuông,…
  + fill = Màu sắc bên trong (vd: của cột hoһc boxplot).
  + color = Đường bên ngoài của cột, boxplot, v.v., hoһc 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).
  + binwidth = 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 hoһc gán cho từng

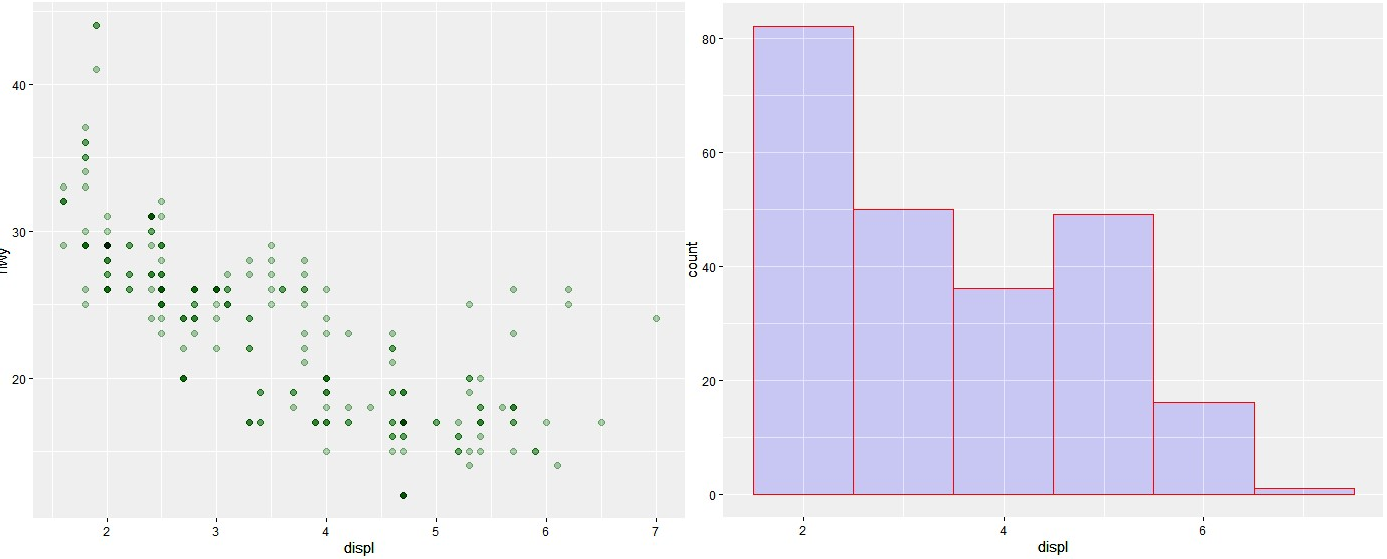
biến của dữ liệu (vd: color = hospital) để hiển thị từng quan sát phụ thuộc vào giá trị của

nó trong biến đó.

* ***Trang tr****í* ***v****ớ****i*** *một giá trị tĩnh*

Nếu muốn yếu tố trang trí cho đối tượng biểu đồ là 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 số mapping = aes() . Các phép gán này có thể ví dụ như: size = 1 hoһc color = "blue".

***V****í* ***d****ụ* ***1.3.*** Xét bộ dữ liệu mpg, với phép gán giá trị tĩnh về màu sắc



# biểu đồ vô hGỚng

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(color = "darkgreen", size = 2, alpha = 0.2)

# biểu đồ phân bố

ggplot(data = mpg, mapping = aes(x = displ))+ geom\_histogram(

binwidth = 1, color = "red", fill = "blue", alpha = 0.1)

# kết quà hi̘n th$

# gán dữ li̘u và các trục

# hàm tạo điểm

# gán dữ li̘u và các trục

# hàm phân bố # độ rộng cột # đGỜng màu

# màu tô bên trong

# độ trong

Hình 1.4: Phép gán tĩnh.

* ***Trang tr****í theo giá trị của từng biến*

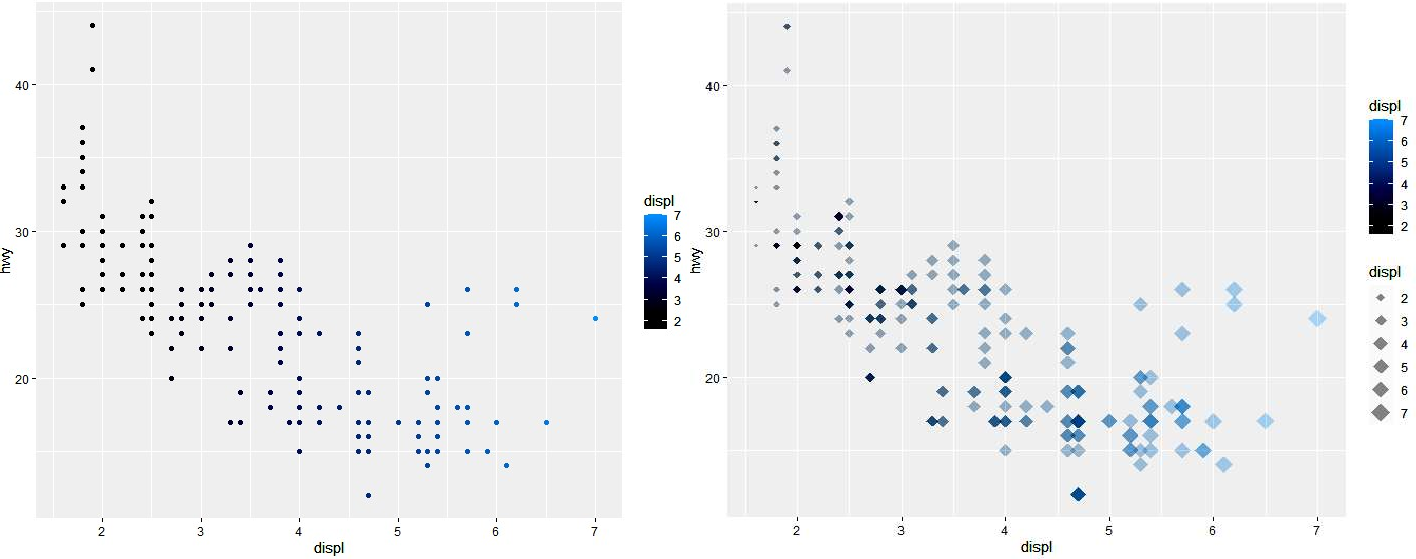
Để thực hiện được điều này, chúng ta gán yếu tố trang trí của biểu đồ với một biến (không

trong dấu ngoһc kép). Điều này phải được thực hiện *bên trong một hàm* mapping = aes()

***V****í* ***d****ụ* ***1.4.*** Xét bộ dữ liệu mpg, với phép trang trí theo biến x = displ bởi màu sắc-color hoһc kích cỡ-size.

# biểu đồ vô hGỚng 1

ggplot(data = mpg, mapping = aes( x = displ, y = hwy, color = displ))+ geom\_point()



# biểu đồ vô hGỚng 2

ggplot(data = mpg, mapping = aes( x = displ, y = hwy,color = displ,size = displ))+ geom\_point(shape = "diamond", alpha = 0.3)

# kết quà hiển th$:

Hình 1.5: Trang trí theo biến.

Trong biều đồ đầu tiên, yếu tố thẩm mỹ color (của mỗi điểm) được gán cho biến displ - và thang đo liên tục được xuất hiện dưới dạng chú thích. Trong biểu đồ thứ hai, hai yếu tố trang trí được gán cho biến displ với hai yếu tố thẩm mỹ là color và size, trong khi shape và alpha được gán cho các giá trị tĩnh bên ngoài đối số mapping = aes().

*Nh¾n xét 1****:*** Các phép gán trục luôn được gán cho các biến trong dữ liệu (không phải cho các giá trị tĩnh) và điều này luôn được thực hiện với mapping = aes(). Điều quan trọng là phải theo dõi các lớp-geom của biểu đồ và các đối tượng thẩm mỹ khi vẽ các biểu đồ phức tạp - ví dụ biểu đồ được cấu thành từ nhiều geoms.

***Nh****¾****n x****é****t 2:*** Việc gán các yếu tố trang trí bên trong đối số mapping = aes() có thể được viết ở một số chỗ trong các lệnh vẽ biểu đồ và thậm chí có thể được viết nhiều lần. Nó có thể được viết trong lệnh ggplot() trên cùng, hoһc cho từng geom riêng lẻ bên dưới. Các kiểu viết bao gồm:

* + Các phép gán được thực hiện ở lệnh ggplot() trên cùng sẽ được mһc định kế thừa ở bất kỳ các geom bên dưới, giống như cách mà x = và y = được kế thừa.
  + Các phép gán được thực hiện trong một geom chỉ áp dụng cho geom đó.
  + Tương tự, data = được chỉ định cho lệnh ggplot() ở trên đầu sẽ áp dụng mһc định cho tất cả các geom bên dưới.

***V****í* ***d****ụ* ***1.6.*** Mỗi lệnh sau sẽ tạo ra cùng một biểu đồ giống nhau:

# Mẩu thứ nhất

ggplot(data = mpg, mapping = aes(x = displ))+ geom\_histogram()

# Mẩu thứ hai

ggplot(data = mpg)+

geom\_histogram(mapping = aes(x = displ)) # Mẩu thứ ba

ggplot()+

geom\_histogram(data = mpg, mapping = aes(x = displ))

* ***Trang tr****í* ***theo nh****óm đối tượ****ng***

*Ví dụ 1.****7.*** Xét lại dữ liệu mpg, để có thông tin chi tiết hơn về dữ liệu, chúng ta có thể sử dụng màu sắc để phân biệt tới biến class hiển thị kiểu dáng của từng chiếc xe. Chúng ta sẽ đһt mapping = aes(color = “class”) khi đó một chú giải tự động xuất hiện . Phép gán này có thể được thực hiện bên trong mapping = aes() ở lệnh ggplot() đầu tiên (và được thừa kế bởi các geom), hoһc nó thể được đһt trong một mapping = aes() riêng biệt bên trong geom. Cả hai cách tiếp cận được trình bày dưới đây:

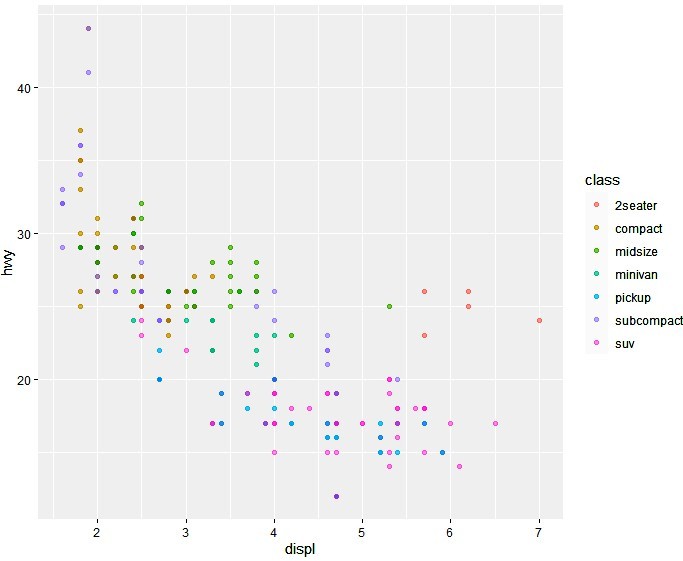
# cách tiếp c̋n 1

ggplot(data = mpg,mapping = aes(x = displ, y = hwy, color = class))+ geom\_point(alpha = 0.5)

# cách tiếp c̋n 2

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(mapping = aes(color = class),alpha = 0.5)

# kết quà hiển th$:



Hình 1.6: Trang trí màu sắc theo nhóm.

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 hoһc 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.

1. *Gán nhãn cho biểu đồ*

Việc đһt 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 đһt bên dưới tiêu đề chính.
  + caption = Chú thích của biểu đồ, mһc định ở góc phải dưới.

***V****í* ***d****ụ* ***1.8.*** Ví dụ dưới đây là biểu đồ chúng ta đã tạo ở Ví dụ 1.7 nhưng có thêm các nhãn:

Bieu\_do1<-ggplot(data = mpg) +

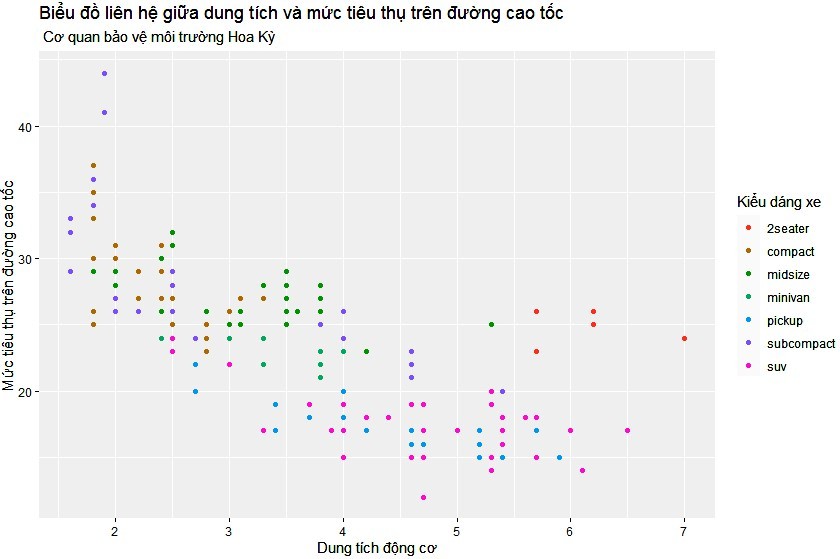
geom\_point(mapping = aes(x = displ, y = hwy,color=class))+

labs(title = "Biểu đồ liên h̘ giữa dung tích và mức tiêu thụ trên đGỜng cao tốc", subtitle = " CƠ quan bào v̘ môi trGỜng Hoa Kỳ",x = "Dung tích động cƠ",y = "Mức

tiêu thụ trên đGỜng cao tốc", color = "Kiểu dáng xe") # in biểu đồ

Bieu\_do1

# Kết quà hiển th$:



Hình 1.7: Gán nhãn cho biểu đồ.

1. *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 mһc định sắn có và căn chỉnh cá nhân đơn lẻ

* *Căn chỉ****nh theo m****¾c đị****nh***

Căn chỉnh theo mһc 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.

* theme\_gray(): Chủ đề ggplot2 đһc 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.
* 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.
* 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.
* 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.
* 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.
* theme\_minimal(): Một chủ đề tối giản không có chú thích nền.
* theme\_classic(): Một chủ đề có giao diện cổ điển với các đường trục x và y và không có đường lưới.
* theme\_void(): Một chủ đề hoàn toàn trống rỗng.
* 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.

***V****í* ***d****ụ* ***1.9.*** Ví dụ dưới đây minh họa một vài căn chỉnh theo chủ đề mһc định:

# cǎn chình theo chù đề Theme classic

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(color = "darkgreen", size = 0.5, alpha = 0.2)+ labs(title = "Theme classic")+

theme\_classic()

# cǎn chình theo chù đề Theme bw

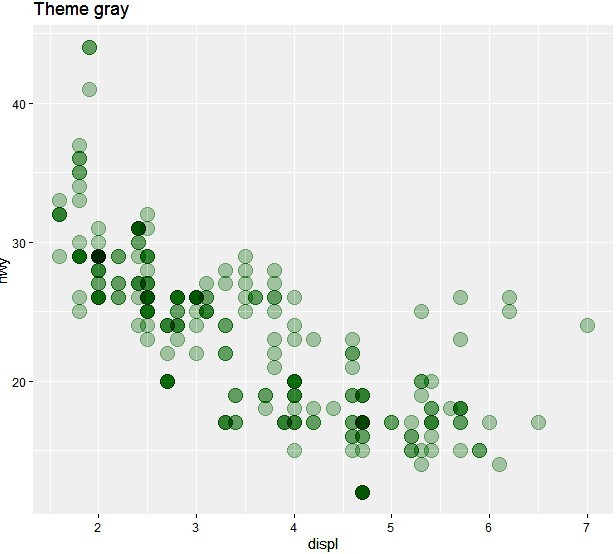
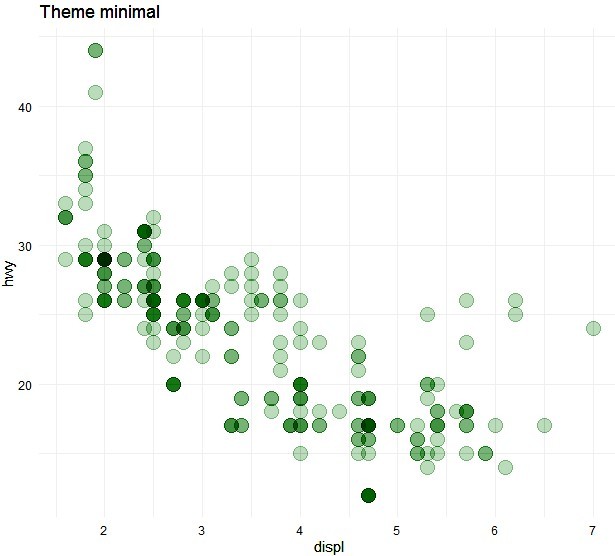
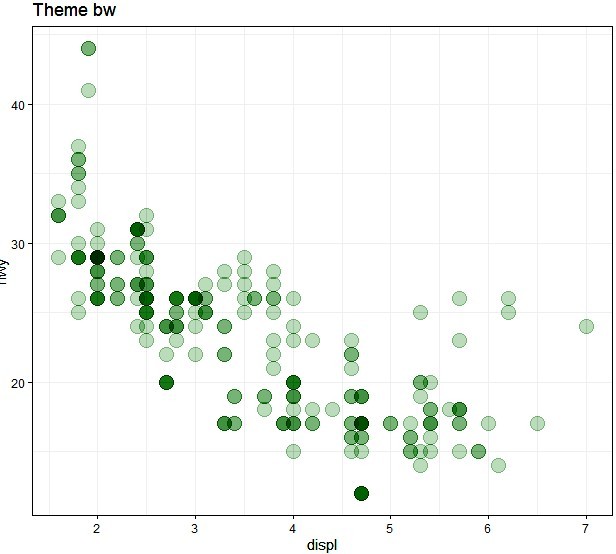
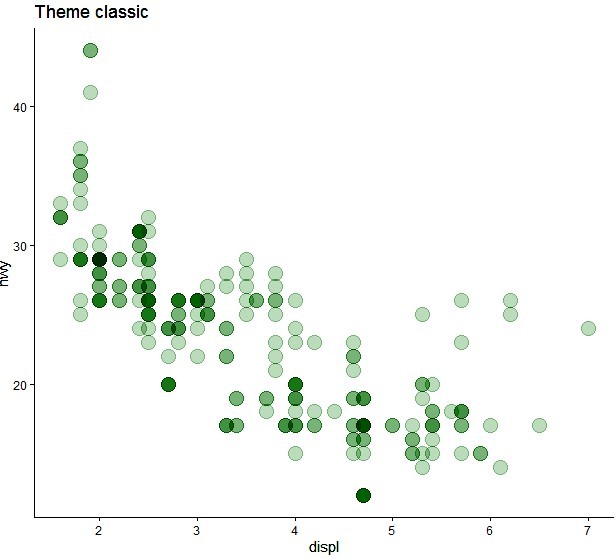
ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(color = "darkgreen", size = 0.5, alpha = 0.2)+ labs(title = "Theme bw")+

theme\_bw()

# cǎn chình theo chù đề Theme minimal

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+

geom\_point(color = "darkgreen", size = 0.5, alpha = 0.2)+ labs(title = "Theme minimal")+



theme\_minimal()

# cǎn chình theo chù đề Theme gray

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(color = "darkgreen", size = 0.5, alpha = 0.2)+ labs(title = "Theme gray")+

theme\_gray()

# kết quà hiển th$:

Hình 1.8: Căn chỉnh theo chủ đề mһc định.

* *Căn chỉ****nh c****á 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:

* + Bên trong hàm theme(), hãy viết tên đối số cho phần tử biểu đồ mà ta muốn chỉnh sửa, chẳng hạn như plot.title =.
  + Cung cấp một hàm element\_() tới đối số.
  + Thường sử dụng nhất là element\_text(), một số khác bao gồm element\_rect() chọn màu nền cho canvas, hoһc element\_blank() để xóa các phần tử biểu đồ.
  + 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().

|  |  |
| --- | --- |
| Đối số | Những điều chỉnh |
| plot.title = element\_text() | Tiêu đề chính |
| plot.subtitle = element\_text() | Tiêu đề phụ |
| plot.caption = element\_text() | Liên quan tới caption (kiểu font, màu sắc, kích cỡ, góc độ, vjust, hjust…) |
| axis.title = element\_text() | Tiêu đề trục (cả trục x và y) (kích cỡ, góc độ, màu sắc…) |
| axis.title.x = element\_text() | Chỉ tiêu đề trục x (sử dụng .y để chỉ áp dụng với trục y) |
| axis.text = element\_text() | Văn bản trên trục (cả trục x và y) |
| axis.text.x = element\_text() | Chỉ văn bản trục x (sử dụng .y để chỉ áp dụng với trục y) |
| axis.ticks = element\_blank() | Loại bỏ ticks của trục |
| axis.line = element\_line() | Đường trục (màu sắc, kích thước, kiểu đường: nét đứt, nét liền mảnh, v.v.) |
| strip.text = element\_text() | Văn bản trong Facet strip (màu sắc, kích thước, góc độ…) |
| strip.background = element\_rect() | facet strip (tô màu, màu sắc, kích thước…) |

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. Hoһc nếu chúng ta muốn xóa một phần tử của biểu đồ, bạn cũng có thể làm điều đó bằng hàm theme(). Chỉ cần đһt element\_blank() tới đối số để nó biến mất hoàn toàn. Đối với chú thích, thiết lập legend.position = "none".

***V****í* ***d****ụ* ***1.10.*** Ví dụ dưới đây sử dụng biểu đồ từ file đã lưu Bieu\_do1 và thêm vào một vài

căn chỉnh theo tùy chọn cá nhân

Bieu\_do1+ theme\_classic()+

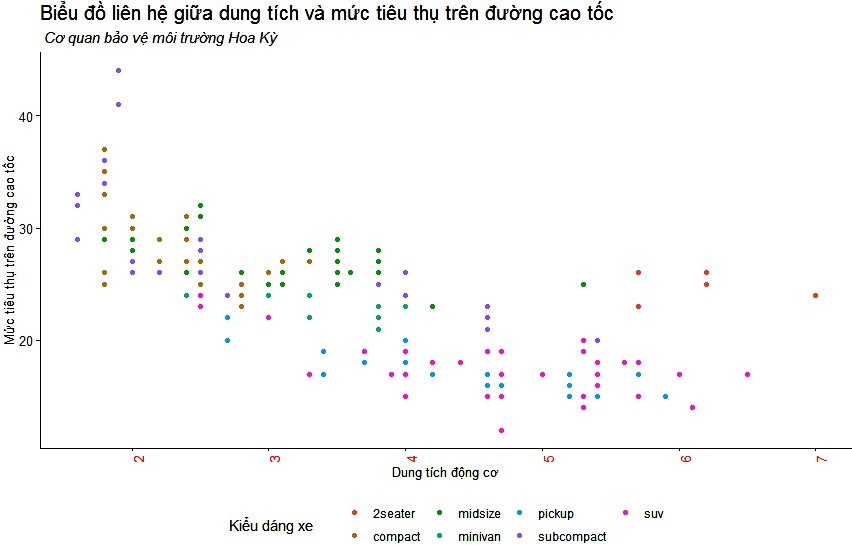
theme(legend.position = "bottom", plot.title = element\_text(size = 15),

plot.caption = element\_text(hjust = 0),

plot.subtitle = element\_text(face = "italic"),

axis.text.x = element\_text(color = "red", size = 10, angle = 90), axis.text.y = element\_text(size = 10),

axis.title = element\_text(size = 10))



Hình 1.9: Căn chỉnh biểu đồ theo tùy chọn.

Giải thích:

* legend.position = là đһc biệt nhất vì nó chỉ chấp nhận các giá trị đơn giản như “bottom”, “top”, “left”, và “right”. các đối số liên quan đến văn bản yêu cầu bạn đһt các chi tiết *bên trong* hàm element\_text().
* Cỡ chư tiêu đề với element\_text(size = 30).
* Căn lề caption với element\_text(hjust = 0) (từ trái qua phải).
* Tiêu đề phụ được in nghiêng với element\_text(face = "italic").

1. *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*) hoһc 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 hoһc 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 hoһc ấn F1).

***V****í* ***d****ụ* ***1.11.*** Ví dụ dưới đây mô tả một vài cách phối màu kết hợp màu bên trong và bên ngoài.

# Phối màu cho điểm và cho các đGỜng

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(color = "purple")+

geom\_vline(xintercept = 4, color = "orange")+ labs(title = " Phối màu cho điểm và cho các đGỜng ")

# Phối màu cho dữ li̘u liên tục

ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(mapping = aes(color = drv))+

labs(title = " Phối màu cho dữ li̘u liên tục ") # Phối màu cho dữ li̘u rỜi rạc

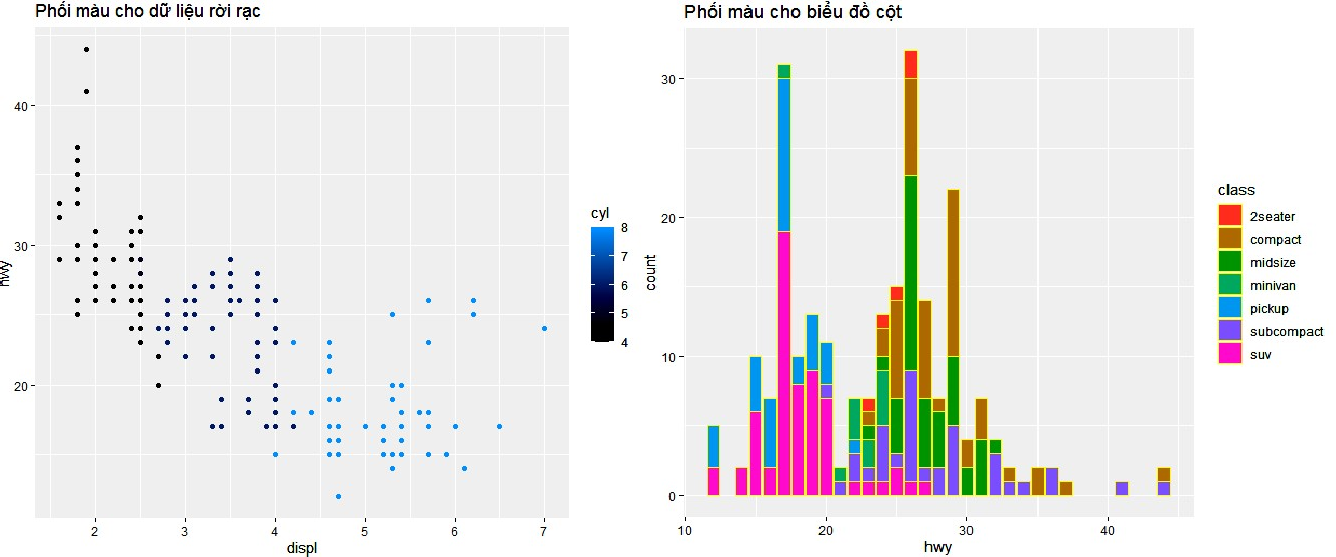
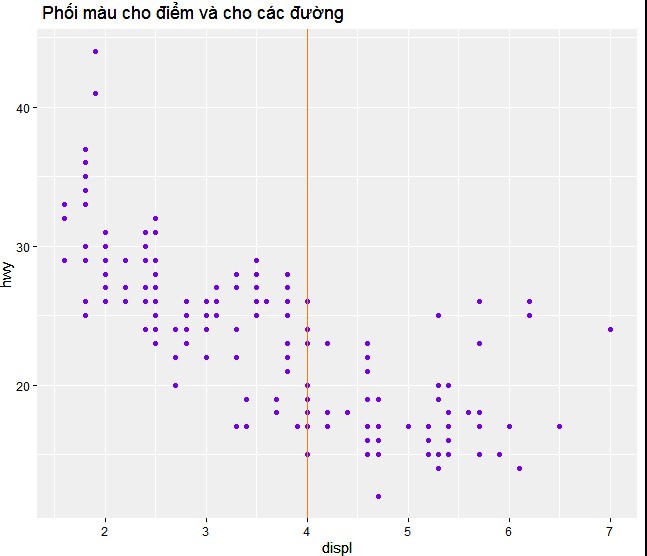
ggplot(data = mpg, mapping = aes(x = displ, y = hwy))+ geom\_point(mapping = aes(color =cyl))+

labs(title = "Phối màu cho dữ li̘u rỜi rạc") # Phối màu cho biểu đồ cột

ggplot(data = mpg, mapping = aes(x = hwy))+ geom\_bar(mapping = aes(fill = class), color = "yellow")+ labs(title = "Phối màu cho biểu đồ cột")

# kết quà hiển th$:

Hình 1.10: Một số cách phối màu.



* *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().

* Phần đầu tiên, scale\_(), là cố định.
* 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\_.
* Phần thứ ba, method, sẽ là một trong số các tùy chọn sau \_discrete(), continuous(), \_date(), \_gradient(), hoһc \_manual(), tùy thuộc vào kiểu dữ liệu củabiến 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, mһc dù cũng 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() (đһt 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\_ hoһc \_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 đһt na.translate = FALSE.

* *Điều chỉnh thủ công*

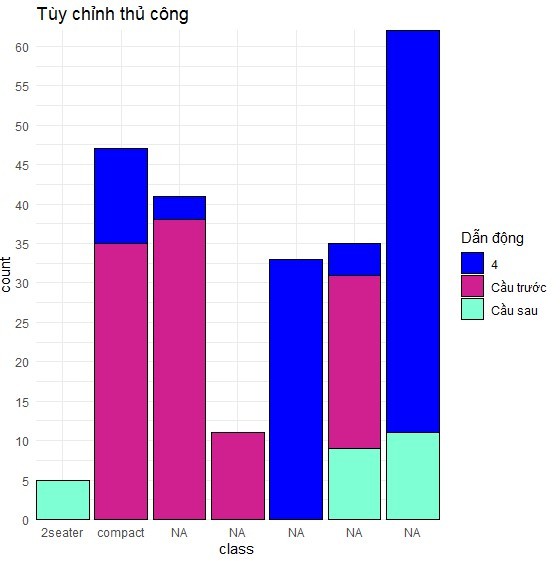
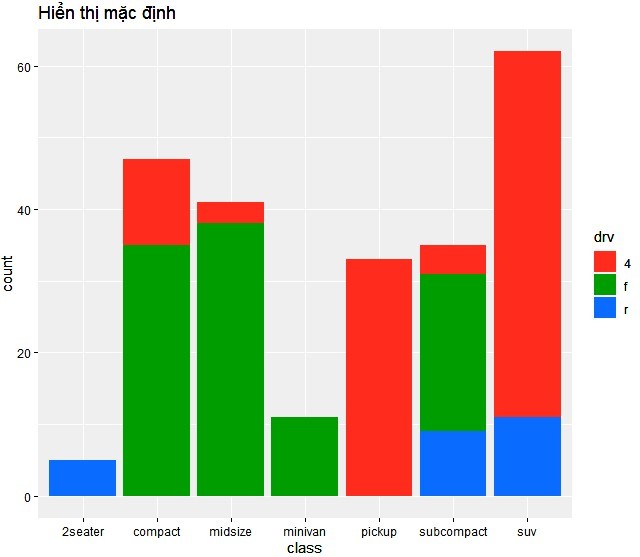
Chúng ta có thể sử dụng các hàm scaling “một cách thủ công” để gán màu sắc như mong muốn.

* + Gán màu cho các giá trị dữ liệu với đối số values =.
  + Cụ thể màu sắc cho giá trị NA với na.value =.
  + Thay đổi cách các giá trị được viết trong chú giải với đối số labels =.
  + Thay đổi tiêu đề chú giải bằng name =.

***V****í* ***d****ụ* ***1.12.*** Dưới đây, chúng ta tạo một biểu đồ cột và hiển thị cách nó xuất hiện theo mһc định, sau đó với ba thang đo được điều chỉnh - thang đo trục y, thang đo trục x, và điều chỉnh thủ công cách tô màu (màu bên trong cột).

# Hiển th$ m̐c đ$nh ggplot(data = mpg)+

geom\_bar(mapping = aes(x = class, fill = drv))+



labs(title = "Hiển th$ m̐c đ$nh") # Hi̘n th$ tùy chình

ggplot(data = mpg)+

geom\_bar(mapping = aes(x = class, fill = drv), color = "black")+ theme\_minimal()+

scale\_y\_continuous(expand = c(0,0),breaks = seq(from = 0,to = 70,by = 5))+ scale\_x\_discrete( expand = c(0,0),drop = FALSE, na.translate = FALSE, labels = c("2seater", "compact"))+

scale\_fill\_manual(

values = c("f" = "violetred","r" = "aquamarine","4" = "blue"),

labels = c("f" = "Cầu trGỚc", "r" = "Cầu sau","4 hGỚng"),name = "Dẩn động", na.value = "grey")+

labs(title = "Tùy chình thù công")

# kết quà hi̘n th$:

Hình 1.11: Điều chỉnh thang đo ghi chú.

* *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 hoһc 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 hoһc 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 =. (xem code trong ví dụ 1.113).

* *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.

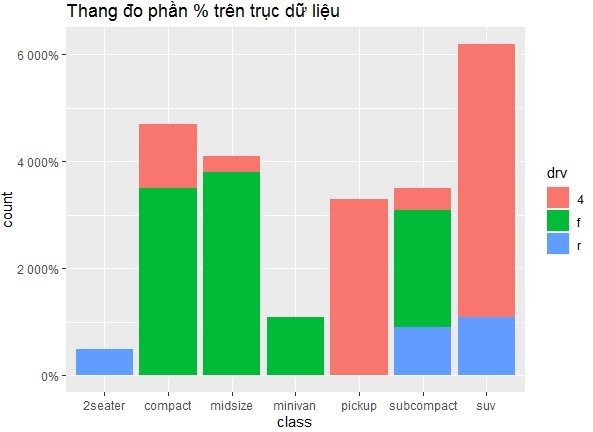
***V****í* ***d****ụ* ***1.13.*** Thêm thang đo phần % vào cột dữ liệu.

# Hiển th$ m̐c đ$nh ggplot(data = mpg)+

geom\_bar(mapping = aes(x = class, fill = drv))+ labs(title = "Thang đo phần % trên trục dữ li̘u")+

scale\_y\_continuous(labels = scales::percent )

# Kết quà hiển th$:



* *Thang đo* ***log***

Hình 1.12: Thang đo hiển thị phần trăm.

Một số dữ liệu khi hiển thị trên biểu đồ có khoảng cách (metric) khá lớn, dẫn tới khó quan sát hoһc 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.

***V****í* ***d****ụ* ***1.14.*** Chúng ta tạo một data frame với các biến region; preparedness\_index và cases\_cumulative như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| plot\_data <- data.frame(  region = c("A", "B", "C", "D", "E", "F", "G", "H", "I"),  preparedness\_index = c(8.8, 7.5, 3.4, 3.6, 2.1, 7.9, 7.0, 5.6, 1.0),  cases\_cumulative = c(15, 45, 80, 20, 21, 7, 51, 30, 1442)) # Kết quà hiển th$: | | | | |
| ## |  | region | preparedness\_index | cases\_cumulative |
| ## | 1 | A | 8.8 | 15 |
| ## | 2 | B | 7.5 | 45 |
| ## | 3 | C | 3.4 | 80 |
| ## | 4 | D | 3.6 | 20 |
| ## | 5 | E | 2.1 | 21 |
| ## | 6 | F | 7.9 | 7 |
| ## | 7 | G | 7.0 | 51 |
| ## | 8 | H | 5.6 | 30 |
| ## | 9 | I | 1.0 | 1442 |

Ta thấy giá trị tích lũy ở vùng “I” lớn hơn đáng kể so với tất cả các vùng khác. Trong những trường hợp như thế này, ta chọn hiển thị trục y bằng thang đo log để có thể thấy sự khác biệt giữa các vùng với ít trường hợp tích lũy hơn.

# Giá tr$ trục y thực tế

preparedness\_plot <- ggplot(data = plot\_data,

mapping = aes(x = preparedness\_index, y = cases\_cumulative))+ geom\_point(size = 2)+

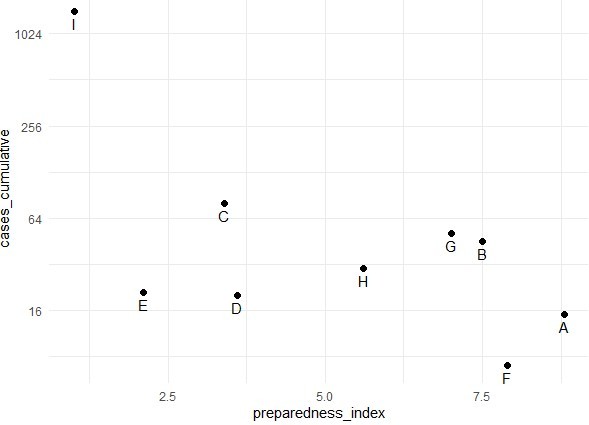
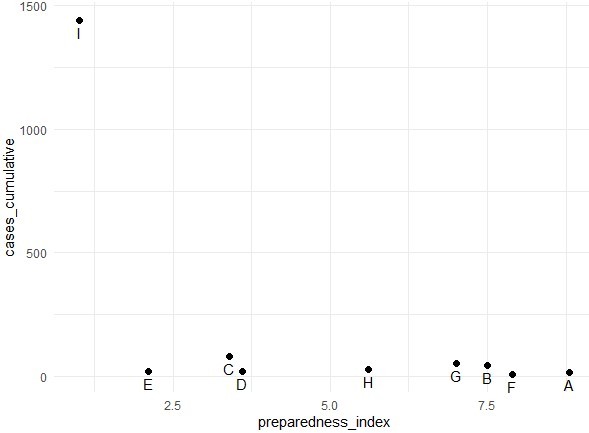
geom\_text( mapping = aes(label = region), vjust = 1.5)+ theme\_minimal()

preparedness\_plot

# Giá tr$ trục y biến đổi theo thang đo log

preparedness\_plot+ scale\_y\_continuous(trans = "log2")

# kết quà hiển th$:



Hình 1.13: Dữ liệu qua phép biến đổi thang đo log.

* *Thang đo Gradient*

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

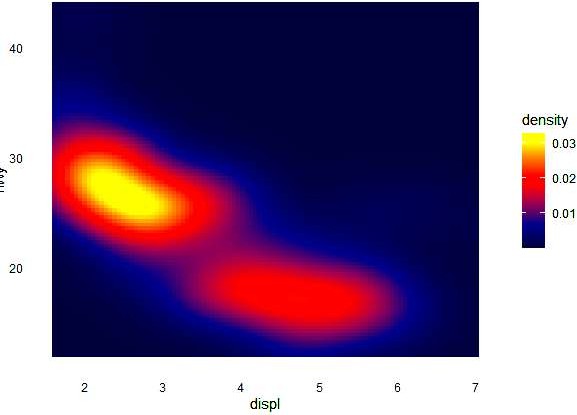
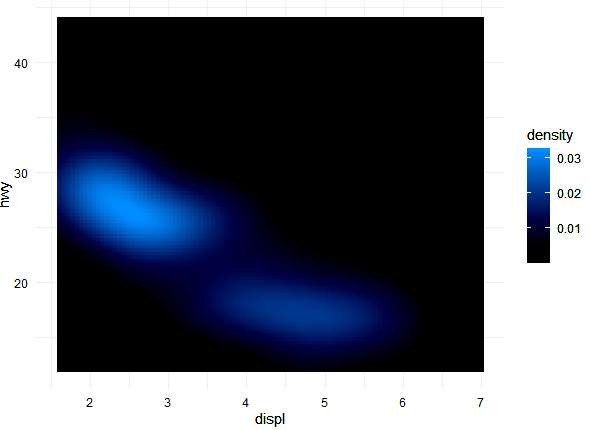
Chúng ta xét ví dụ sau đây với biểu đồ nhiệt, lần lượt điều chỉnh thang đo màu gradien.

***V****í* ***d****ụ* ***1.15.*** Xét biểu đồ nhiệt với màu gốc và màu hiệu ứng “plasma”

trans\_matrix <- ggplot( data = mpg,mapping = aes(x = displ, y = hwy))+ stat\_density2d(

geom = "raster",mapping = aes(fill = after\_stat(density)),contour = FALSE)+ theme\_minimal()

trans\_matrix



trans\_matrix + scale\_fill\_viridis\_c(option = "plasma") # kết quà hiển th$:

Hình 1.14: Biểu đồ dữ liệu với hiệu ứng gốc và “plasma”.

Tiếp theo, chúng ta sẽ so sánh kết quả khi điều chỉnh các điểm ngắt của thang đo qua một số hàm sau:

* scale\_fill\_gradient() nhận hai màu (cao/thấp).
* scale\_fill\_gradientn() nhận một vectơ có độ dài màu bất kỳ tới values = (các giá trị trung gian sẽ được nội suy).
* Sử dụng scales::rescale() để điều chỉnh cách định vị màu sắc dọc theo gradient; nó sẽ cân chỉnh lại vectơ vị trí nằm giữa 0 và 1.

trans\_matrix +

scale\_fill\_gradient(low ="aquamarine",high ="purple",

na.value ="grey", name = "Density")+ labs(title = "Phân hai màu cao/thấp")

# Thang màu 3 loại trans\_matrix +

scale\_fill\_gradientn( colors = c("blue", "yellow","red") )+ labs(title = "Thang màu 3 loại: thấp, trung bình , cao")

# Điều chình màu theo tỷ l̘

trans\_matrix + scale\_fill\_gradientn(

colors = c("blue", "yellow","red", "black"),

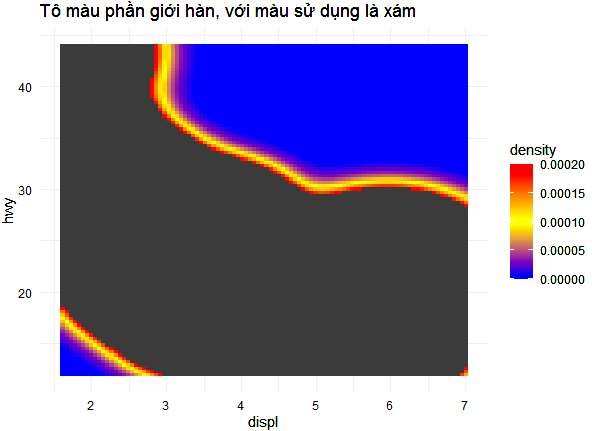
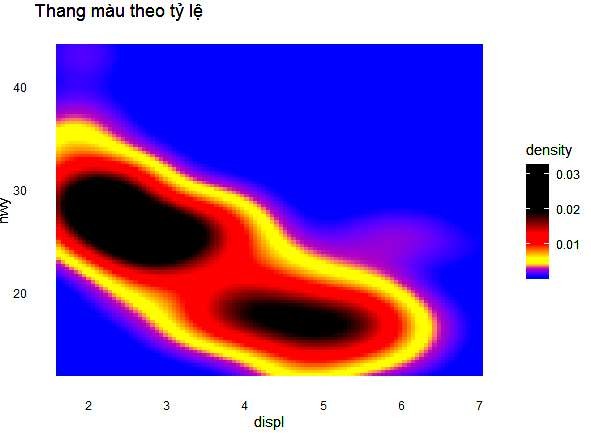
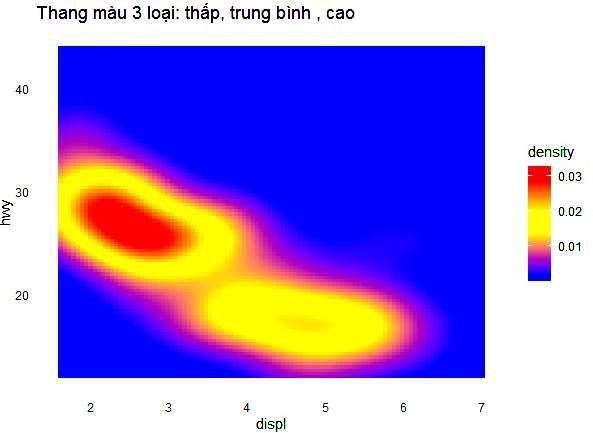
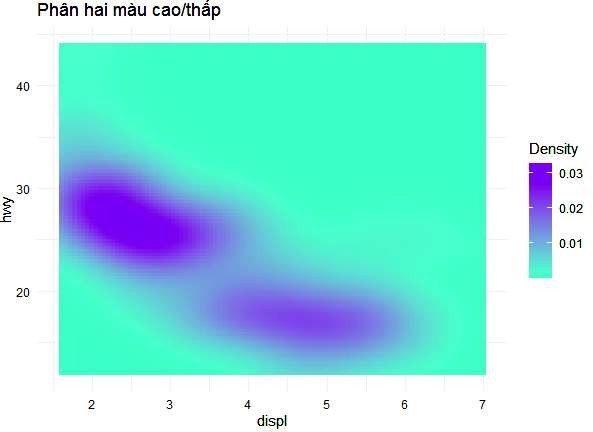
values = scales::rescale(c(0, 0.05, 0.07, 0.10, 0.15, 0.20, 0.3, 0.5)))+ labs(title = "Thang màu theo tỷ l̘")

# Sừ dụng các giá tr$ giỚi hạn để tô màu trans\_matrix +

scale\_fill\_gradientn(

colors = c("blue", "yellow","red"), limits = c(0, 0.0002))+ labs(title = "Tô màu phần giỚi hàn, vỚi màu sừ dụng là xám ")

# kết quà hi̘n th$



Hình 1.15: Một số tùy chỉnh thang màu qua các hàm gradian.

1. *Lưu trũ, chỉnh sửa và xuất bản biểu đồ*
   * *Lưu biểu đồ*

Mһc 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 đһt 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.

***V****í* ***d****ụ* ***1.16.*** lệnh lưu biểu đồ hình 1.1 vào biến có tên Bieu\_do1

Bieu\_do1<-ggplot(data = mpg) +

geom\_point(mapping = aes(x = displ, y = hwy,color=class)) # in biểu đồ

Bieu\_do1

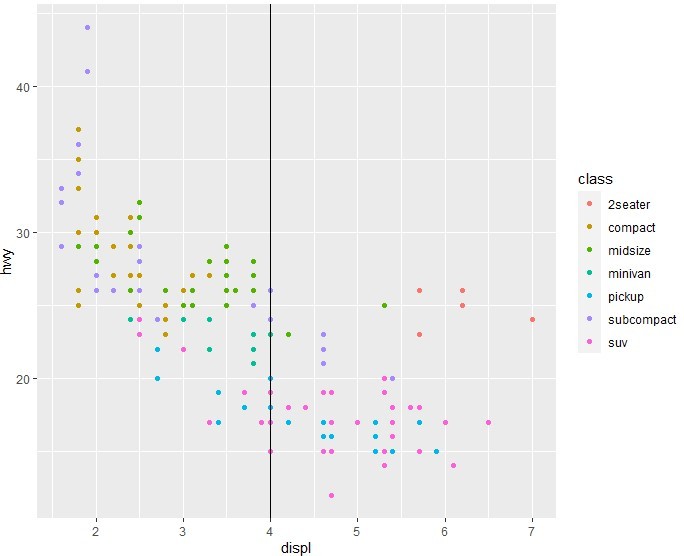
* + *Chỉnh sửa biều đồ đã lưu*

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 hoһc chỉnh sửa bắt đầu bằng tên của nó. Chúng ta không cần phải lһp lại tất cả các lệnh đã tạo ra biểu đồ ban đầu.

*Ví dụ* ***1.17***. Ta chỉnh sửa đổi biểu đồ Bieu\_do1 đã được lưu ở Ví dụ 1.16 ở bên trên, thêm một trục dọc tại dung tích động cơ bằng 4, chúng ta chỉ cần thêm dấu + và bắt đầu thêm các lớp bổ sung vào biểu đồ.

Bieu\_do1+ geom\_vline(xintercept = 4)

# kết quà hiển th$:



Hình 1.16: Chỉnh sửa biểu đồ đã lưu trữ.

* + *Xuất bản biểu đồ*

Việc xuất bản biểu đồ được thực hiện dễ dàng với hàm ggsave() của package *ggplot2*

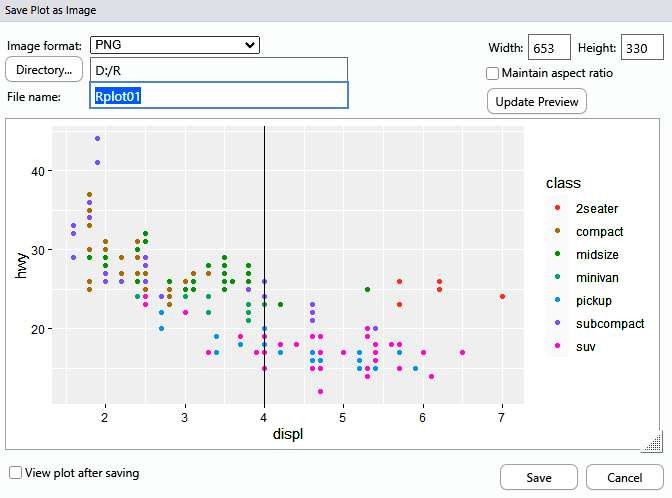
hoһc chức năng Export trong Rstudio.

* 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 dụ: 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, hoһc 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 giâỉ của biểu đồ (vd: dpi = 300). Xem hướng dẫn chi tiết về hàm bằng cách gõ ?ggsave trong Rstudio.

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



Hình 1.17: 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. PHÂN TÍCH THỐNG KÊ

GIỚI THIỆU CHUNG

Do đặc thù khác nhau của việc thu thập và xử lý các số liệu thống kê nên trong khuôn khổ của cuốn sách này chúng tôi mong muốn tập trung vào những vấn đề về các số liệu thống kê phục vụ cho các nghiên cứu về lĩnh vực kinh tế - xã hội.Khái niệm phân tích số liệu thống kê: Là sự kết hợp giữa  
thống kê, sự tư duy và hiểu biết các vấn đề kinh tế.

Yêu cầu: Để có thể nắm vững kiến thức của môn học này đòi hỏi người học phải có kiến thức sâu về thống kê, về kinh tế cũng như những hiểu biết thực tế của vấn đề nghiên cứu. Ngoài ra, cần phải có kiến thức về tin học và các công cụ lượng hoá khác để kết hợp trong nghiên cứu.

Trước khi bước vào nội dung của chương trình thứ nhất chúng tôi muốn trao đổi sơ lược với các độc giả về tổng quát tiến hành một nghiên cứu trong các vấn đề thuộc về kinh tế - xã hội.

Trong khi tiến hành các nghiên cứu về kinh tế - xã hội có gì khác so với các vấn đề thuộc về khoa học tự nhiên: điều khác cơ bản đó là đối tượng nghiên cứu trong các nghiên cứu kinh tế - xã hội thường là con người hoặc là liên quan đến con người, các mối liên hệ với con người. Ngoài ra, nó còn khác nhau ở cách  
thức tiến hành, khả năng áp dụng và thời gian cho kết quả, phạm vi tác động v.v...

Thiết kế một nghiên cứu về kinh tế - xã hội cần phải làm những gì? Dưới đây là một đề cương sơ bộ hướng dẫn cho các nghiên cứu thuộc về lĩnh vực khoa học kinh tế - xã hội. Nó sẽ được cụ thể hoá cho từng chương trình nghiên cứu cụ thế.

**2.1 Vấn đề đặt ra**

A. Trình bày một cách rõ ràng, ngắn gọn về vấn đề đặt ra, với việc xác định khái niệm cần thiết như thế nào.

B. Chỉ ra vấn đề là sự giới hạn về ranh giới để giải quyết hoặc kiểm tra vấn đề.

C. Mô tả sự cần thiết và ý nghĩa của vấn đề liên quan đến  
một trong những chi tiết sau:

1. Thời gian.
2. Liên quan đến vấn đề thực tế.
3. Liên quan đến tổng thể rộng lớn hơn.
4. Liên quan đến sự tác động hoặc phản ảnh đến tổng thể.
5. Làm thoả mãn khoảng cách của một nghiên cứu.
6. Cho phép suy rộng ra các hoạt động xã hội hoặc các  
   nguyên lý cơ bản.
7. Làm rõ các khái niệm, mối quan hệ và sự quan trong.
8. Tìm hiểu phạm vi thực tế của vấn đề trong thực tế.
9. Có thể tạo ra hoặc phát triển những công cụ quan trọng  
   cho việc quan sát và phân tích thông tin.
10. Cung cấp cơ hội và khả năng thu thập thông tin trong  
    thực trạng của việc hạn chế về thời gian.
11. Trình bày khả năng có thể giải thích hoặc phân tích kết  
    quả một cách tốt nhất, có nhiều thông tin nhất dựa trên cơ sở các  
    kỹ thuật phân tích đã có.

**2.2 Cơ sở lý luận**

A. Trình bày mối quan hệ của vấn đề đến cơ sở lý luận.

B. Sự liên quan của vấn đề nghiên cứu tới các nghiên cứu  
trước đây.

C. Trình bày các giả cứ lý luận liên quan.

**2.3 Giả thuyết**

1. Làm rõ các giả thuyết lựa chọn cho việc kiểm định.
2. Thể hiện mức độ ý nghĩa của kiểm định các giả thuyết tới sự tiến bộ của nghiên cứu và lý luận.
3. Định nghĩa các khái niệm hoặc các biến sử dụng (tốt nhất nên ở dạng quan hệ phụ thuộc).

*Ví dụ: Thu nhập là phần còn lại của doanh thu sau khi trừ đi chi phí trước khi tính công lao động.*

1. Các biến độc lập (biến giải thích) và biến phụ thuộc (biến được giải thích) giữa chúng nên được phân biệt rõ.
2. Tỷ lệ trên đó các biến được xác định và đo đạc (định lượng, bán định lượng hay định tính) cần được cụ thể.
3. Miêu tả những lỗi có thể mắc phải và hậu quả của nó.
4. Chú thích các lỗi nghiêm trọng.

**2.4 Thiết kế một thí nghiệm hay cuộc điều tra**

1. Trình bày ý tưởng những thiết kế với những quan tâm cụ thể trong việc đáp ứng tính phức tạp của các biến.
2. Mô tả việc lựa chọn một thiết kế để tiến hành.
3. Mô tả các tác nhân kích thích, chủ đề, môi trường và câu hỏi của các mục tiêu, các sự kiện, nhu cầu cần thiết về vật lực.
4. Mô tả làm thế nào để điều khiển được tính phức tạp của các biến.
5. Cụ thể các công cụ dùng để kiểm định thống kê bao gồm cả các bảng giá định cho mỗi một kiểm định.

Trong đó, cần cụ thể mức độ tin cậy mong muốn.

**2.5 Quá trình chọn mẫu**

1. Mô tả mẫu được lựa chọn trong thí nghiệm hoặc điều tra.
2. Cụ thể tổng thể có liên quan đến giả thuyết nghiên cứu.
3. Giải thích sự xác định của số lượng và kiểu loại mẫu.
4. Cụ thể hoá phương pháp lựa chọn mẫu.
5. Cụ thể hoá mối quan hệ tương đối của sai số ngẫu nhiên và phi ngẫu nhiên.
6. Ước lượng chi phí tương đối của các cỡ và kiểu lấy mẫu khác nhau phù hợp với lý thuyết.

**2.6 Phương pháp và cách thức điều tra**

1. Mô tả thước đo của các biến định lượng chỉ ra tính tin cậy và hợp lệ của chúng. Mô tả phương tiện để xác định cho các biến định tính.
2. Các mục bao gồm trong bảng câu hỏi điều tra.
3. Số lượng câu hỏi chó thể phỏng vấn người được hỏi.
4. Thời gian có thể cho cuộc phỏng vấn.
5. Lịch trình tiến hành trong thời gian cụ thể.
6. Những kết quả đánh giá, kiểm định trước.
7. Các mục bao gồm trong quá trình điều tra.
8. Các phương tiện thu thập thông tin.

Ví dụ: Phỏng vấn trực tiếp, hoặc một phần bằng thư, điện thoại hay các phương tiện khác.

1. Các đặc trưng riêng mà người điều tra viên phải có hoặc cần phải tập huấn cho họ.
2. Mô tả kết quả sử dụng từ các nghiên cứu đại diện hoặc điều tra thử. Trong đó nêu rõ sự quan trọng, các phương tiên để xử lý tình trạng thông tin kém giá trị, bị loại bỏ hoặc do lỗi người được hỏi.

**2.7 Hướng dẫn trong quá trình tiến hành**

A. Chuẩn bị một hướng dẫn trong toàn bộ quá trình tiến hành nghiên cứu, trong đó trình bày cụ thể thời gian và ước tính chi phí.

1. Kế hoạch.
2. Địa điểm nghiên cứu và kiểm tra trước.
3. Lựa chọn mẫu.
4. Chuẩn bị các trang thiết bị vật chất cho điều tra.
5. Lựa chọn điều tra viên và tiến hành tập huấn.
6. Kế hoạch triển khai thực địa.
7. Chỉnh sửa lại kế hoạch.
8. Thu thập thông tin.
9. Phân tích thông tin (số liệu).
10. Chuẩn bị báo cáo kết quả nghiên cứu.

**2.8 Phân tích số liệu**

Cụ thể các phương pháp dùng trong phân tích:

1. Sử dụng bảng biểu, các công cụ tính toán, cách thức phân loại, máy tính v.v...
2. Sử dụng các kỹ thuật đồ hoạ.
3. Cụ thể các loại bảng biểu sẽ thiết kế.

**2.9 In quyển hoặc báo cáo kết quả**

1. Kết quả được viết và in ấn theo yêu cầu của đơn vị đào tạo hoặc nghiên cứu.
2. Lựa chọn kết quả viết báo cáo cho các tạp chí khoa học.

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

### **1. PHƯƠNG PHÁP PHÂN TÍCH THĂM DÒ DỮ LIỆU**

### 1.1 GIỚI THIỆU

Phân tích thăm dò dữ liệu (Exploratory Data Analysis - EDA) là một giai đoạn quan trọng trong quy trình phân tích dữ liệu, giúp hiểu sâu hơn về các đặc điểm, mối quan hệ và xu hướng của dữ liệu. Trong lĩnh vực xăng dầu, EDA hỗ trợ việc nhận diện các yếu tố ảnh hưởng đến giá cả, sản lượng và tiêu thụ. Báo cáo này sẽ trình bày chi tiết các phương pháp và công cụ EDA, cùng với các ví dụ minh họa cụ thể về dữ liệu xăng dầu.

### 1.2 THỐNG KÊ MÔ TẢ (DESCRIPTIVE STATISTICS)

Ảnh có chứa văn bản, ảnh chụp màn hình, thiết kế đồ họa, Đồ họa

Mô tả được tạo tự động

Thống kê mô tả cung cấp cái nhìn tổng quan về dữ liệu qua các chỉ số cơ bản như:

* Trung bình (Mean): Giá trị trung bình của dữ liệu, cho biết mức độ trung bình của giá xăng dầu.
* Trung vị (Median): Giá trị giữa của dữ liệu, giúp hiểu phân phối dữ liệu.
* Độ lệch chuẩn (Standard Deviation): Đo lường mức độ biến động của dữ liệu.
* Giá trị cực đại (Max) và cực tiểu (Min): Xác định biên độ của dữ liệu.

Ví dụ: Dữ liệu giá xăng dầu:

Thống kê mô tả cho RON 95-V:

Vùng 1: Mean (trung bình): 24.21 Độ lệch chuẩn (standard deviation): 0.91 Giá trị tối thiểu: 22.72 Giá trị tối đa: 26.27

Vùng 2: Mean: 24.63 Độ lệch chuẩn: 0.94 Giá trị tối thiểu: 22.36 Giá trị tối đa: 26.79

Thống kê mô tả cho RON 95-III:

Vùng 1: Mean: 23.50 Độ lệch chuẩn: 1.11 Giá trị tối thiểu: 21.93 Giá trị tối đa: 25.74

Vùng 2: Mean: 23.83 Độ lệch chuẩn: 1.15 Giá trị tối thiểu: 21.28 Giá trị tối đa: 26.25

Thống kê mô tả cho E5 RON 92:

Vùng 1: Mean: 22.63 Độ lệch chuẩn: 1.11 Giá trị tối thiểu: 20.01 Giá trị tối đa: 24.84

Vùng 2: Mean: 22.84 Độ lệch chuẩn: 1.15 Giá trị tối thiểu: 20.81 Giá trị tối đa: 24.93

Thống kê mô tả cho Diezen 0,001S-V:

Vùng 1: Mean: 21.70 Độ lệch chuẩn: 1.18 Giá trị tối thiểu: 18.84 Giá trị tối đa: 24.45

Vùng 2: Mean: 21.96 Độ lệch chuẩn: 1.17 Giá trị tối thiểu: 19.21 Giá trị tối đa: 24.93

Thống kê mô tả cho Diezen 0,05S-ll\_V1:

Vùng 1: Mean: 20.52 Độ lệch chuẩn: 1.18 Giá trị tối thiểu: 17.95 Giá trị tối đa: 23.59

Vùng 2: Mean: 20.99 Độ lệch chuẩn: 1.19 Giá trị tối thiểu: 18.00 Giá trị tối đa: 24.06

Thống kê mô tả cho Dầu hỏa 2 - K\_V1:

Vùng 1: Mean: 20.99 Độ lệch chuẩn: 1.07 Giá trị tối thiểu: 17.77 Giá trị tối đa: 23.81

Vùng 2: Mean: 21.32 Độ lệch chuẩn: 1.10 Giá trị tối thiểu: 18.12 Giá trị tối đa: 24.28

Thống kê mô tả cho Mazút N°2B (3,5S)\_V1:

Vùng 1: Mean: 16.08 Độ lệch chuẩn: 0.98 Giá trị tối thiểu: 14.42 Giá trị tối đa: 17.84

Vùng 2: Mean: 16.10 Độ lệch chuẩn: 0.96 Giá trị tối thiểu: 14.70 Giá trị tối đa: 18.19

Thống kê mô tả cho Mazút 180cst - 0,5S (RMG)\_V1:

Vùng 1: Mean: 20.48 Độ lệch chuẩn: 0.92 Giá trị tối thiểu: 18.29 Giá trị tối đa: 22.75

Vùng 2: Mean: 20.76 Độ lệch chuẩn: 0.90 Giá trị tối thiểu: 18.61 Giá trị tối đa: 22.74

### 1.3 TRỰC QUAN HÓA DỮ LIỆU (DATA VISUALIZATION)

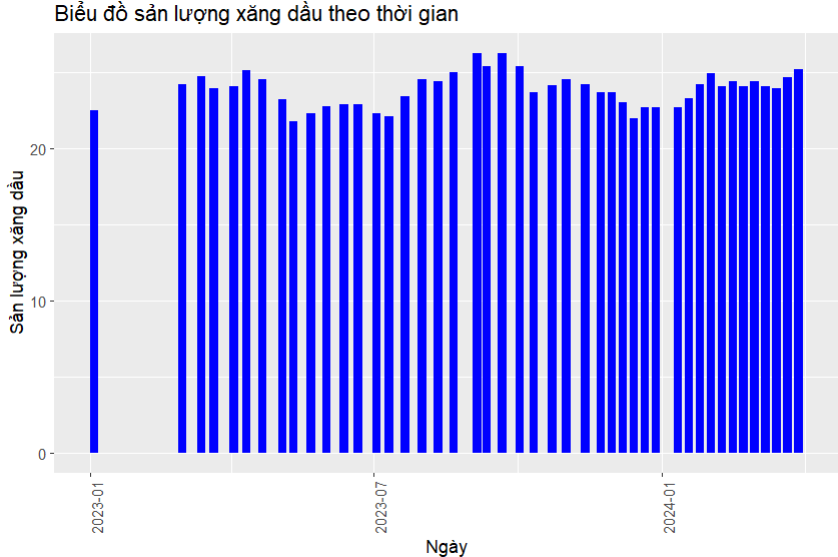
Trực quan hóa dữ liệu giúp dễ dàng nhận diện các mẫu và xu hướng. Các công cụ phổ biến bao gồm:

* Biểu đồ đường (Line Chart): Hiển thị sự thay đổi của giá xăng dầu theo thời gian.

A graph with blue lines

Description automatically generated

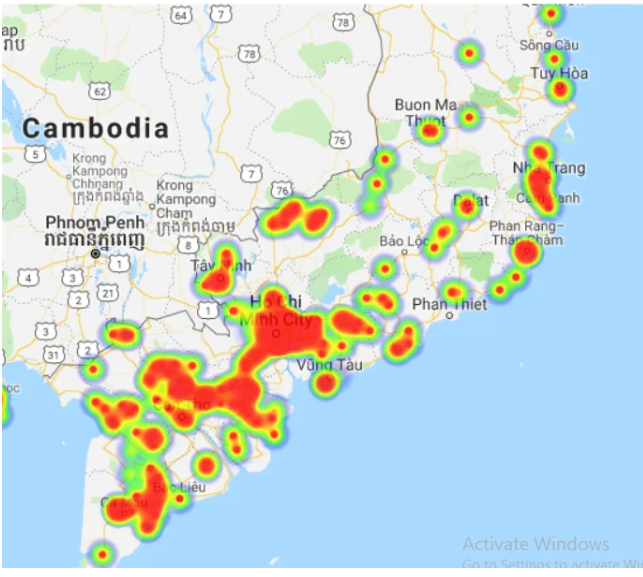
* Biểu đồ thanh (Bar Chart): So sánh sản lượng xăng dầu giữa các thời kỳ.



* Biểu đồ phân tán (Scatter Plot): Tìm hiểu mối quan hệ giữa các biến số như giá dầu thô và giá xăng bán lẻ.

A graph with black dots

Description automatically generated

* Bản đồ nhiệt (Heat Map): Hiển thị mật độ hoặc cường độ tiêu thụ xăng dầu theo vùng.

### 1.4 PHÂN TÍCH TƯƠNG QUAN (CORRELATION ANALYSIS)

Phân tích tương quan giúp xác định mối quan hệ giữa các biến số trong dữ liệu. Hệ số tương quan (Correlation Coefficient) nằm trong khoảng từ -1 đến 1, cho biết mức độ và hướng của mối quan hệ.

Ví dụ: NHẬN XÉT SỰ TƯƠNG QUAN CỦA XĂNG DẦU PETROLIMEX VÙNG I

RON 95\_V\_V1 có mối tương quan rất mạnh với RON\_95\_III\_V1 và E5\_RON\_92\_II\_V1, với các hệ số tương quan lần lượt là 0.98, 0,97

RON 95\_V\_V1 cũng có mối tương quan mạnh với Diezen\_S\_I VI và Diezen\_S II VI, với các hệ số tương quan lần lượt là 0.82 và 0.81.

RON 95\_V\_V1 cũng có mối tương quan mạnh với Dauhoa 2KVI và Mazut\_NZB VI, với các hệ số tương quan lần lượt là 0.73 và 0.70.

Tuy nhiên, RON 95\_V\_V1 chỉ có mối tương quan trung bình với Mazu\_180cSt\_RMG VI với hệ số tương quan là 0.47.

Điều này có nghĩa là khi giá trị của RON 95\_V\_V1 thay đổi, hầu hết các biến số khác cũng sẽ thay đổi theo cùng một hướng.

E5\_RON\_92\_11\_V1 và RON\_95\_V\_V1 ,RON\_95\_111\_V1,Diezen\_SV\_V1,Diezen\_S\_II\_V1,Dauhoa\_2\_KV1,Mazut\_N2B\_có mối tương quan rất manh với nhau với hệ số tương quan lần lượt là :0,97, 0,98, 0.80, 0.82, 0.76, 0,78

Điều này có nghĩa là khi E5\_RON\_92\_11\_V1 thay đổi thì các biến số khác sẽ thay đổi cùng hướng

Tuy nhiên,E5\_RON 92\_11\_V1 chỉ có mối tương quan trung bình với Mazu\_180cSt\_RMG VI với hệ số tương quan là 0.58

Diezen\_SV\_V1 và RON\_95\_V\_V1 ,RON\_95\_111\_V1,Diezen\_S\_II\_V1,Dauhoa\_2\_KV1,Mazut\_N2B\_có mối tương quan rất manh với nhau với hệ số tương quan lần lượt là : 0.82, 0.84, 0.80, 0.98, 0.95, 0.77 đều là mối tương quan dương

Tuy nhiên, Diezen\_SV\_V1 chỉ có mối tương quan yếu với Mazu\_180cSt\_RMG VI với hệ số tương quan là 0.43

### 1.5 PHÁT HIỆN DỮ LIỆU NGOẠI LỆ (OUTLIER DETECTION)

Dữ liệu ngoại lệ có thể ảnh hưởng đến kết quả phân tích và cần được xem xét kỹ lưỡng. Các phương pháp phát hiện bao gồm:

* A blue line with black lines

  Description automatically generatedBiểu đồ hộp (Box Plot): Hiển thị các giá trị ngoại lệ qua các điểm nằm ngoài phần tử hộp.
* Phân tích giá trị z (Z-score Analysis): Xác định các điểm dữ liệu nằm xa giá trị trung bình.

|  |  |  |  |
| --- | --- | --- | --- |
| STT | DATE | RON\_95\_V\_V1 | Z\_score |
| 1 | 28/3/2024 | 25.22 | 1.2169534 |
| 2 | 21/3/2024 | 24.69 | 0.736577 |
| 3 | 14/3/2024 | 23.95 | 0.0658629 |
| 4 | 7/3/2024 | 24.06 | 0.1655637 |
| 5 | 29/2/2024 | 24.43 | 0.5009207 |
| 6 | 22/2/2024 | 24.1 | 0.2018185 |
| 7 | 15/2/2024 | 24.42 | 0.491857 |
| 8 | 8/2/2024 | 24.05 | 0.1565 |
| 9 | 1/2/2024 | 24.95 | 0.9722334 |
| 10 | 25/1/2024 | 24.19 | 0.2833918 |
| 11 | 18/1/2024 | 23.27 | -0.550469 |
| 12 | 11/1/2024 | 22.72 | -1.048973 |
| 13 | 4/1/2023 | 22.47 | -1.275565 |
| 14 | 28/12/2023 | 22.7 | -1.0671 |
| 15 | 21/12/2023 | 22.7 | -1.0671 |
| 16 | 14/12/2023 | 21.96 | -1.737814 |
| 17 | 7/12/2023 | 23.01 | -0.786125 |
| 18 | 30/11/2023 | 23.68 | -0.178857 |
| 19 | 23/11/2023 | 23.71 | -0.151666 |
| 20 | 13/11/2023 | 24.22 | 0.3105829 |
| 21 | 1/11/2023 | 24.56 | 0.6187489 |
| 22 | 23/10/2023 | 24.15 | 0.247137 |
| 23 | 11/10/2023 | 23.68 | -0.178857 |
| 24 | 2/10/2023 | 25.37 | 1.3529089 |
| 25 | 21/09/2023 | 26.27 | 2.1686423 |
| 26 | 11/9/2023 | 25.4 | 1.3801001 |
| 27 | 5/9/2023 | 26.27 | 2.1686423 |
| 28 | 21/08/2023 | 25.03 | 1.044743 |
| 29 | 11/8/2023 | 24.42 | 0.491857 |
| 30 | 1/8/2023 | 24.56 | 0.6187489 |
| 31 | 21/07/2023 | 23.39 | -0.441705 |
| 32 | 11/7/2023 | 22.09 | -1.619986 |
| 33 | 3/7/2023 | 22.3 | -1.429648 |
| 34 | 21/6/2023 | 22.89 | -0.89489 |
| 35 | 12/6/2023 | 22.89 | -0.89489 |
| 36 | 1/6/2023 | 22.79 | -0.985527 |
| 37 | 22/5/2023 | 22.27 | -1.456839 |
| 38 | 11/5/2023 | 21.78 | -1.900961 |
| 39 | 4/5/2023 | 23.22 | -0.595788 |
| 40 | 21/4/2023 | 24.53 | 0.5915578 |
| 41 | 11/4/2023 | 25.14 | 1.1444437 |
| 42 | 3/4/2023 | 24.05 | 0.1565 |
| 43 | 21/03/2023 | 23.96 | 0.0749266 |
| 44 | 13/03/2023 | 24.74 | 0.7818956 |
| 45 | 1/3/2023 | 24.23 | 0.3196466 |

Ví dụ: Biểu đồ hộp cho thấy một số giá trị ngoại lệ trong dữ liệu giá xăng dầu, có thể do các sự kiện bất thường như thiên tai hoặc khủng hoảng kinh tế.

### 1.6 PHÂN TÍCH PHÂN PHỐI (DISTRIBUTION ANALYSIS)

Phân tích phân phối giúp hiểu rõ cách dữ liệu được phân bố, từ đó xác định các đặc điểm như lệch hay đối xứng của dữ liệu. Các công cụ bao gồm:

* Biểu đồ histogram: Hiển thị tần suất của các giá trị dữ liệu.

A graph with blue squares

Description automatically generated

* A blue graph with numbers

  Description automatically generatedBiểu đồ đường cong phân phối: Cho thấy hình dạng của phân phối dữ liệu

Ví dụ: Biểu đồ histogram của giá xăng dầu cho thấy dữ liệu có phân phối gần với phân phối chuẩn, nhưng có một số giá trị lệch phải.

### **2. CÔNG CỤ SỬ DỤNG TRONG EDA**

### 2.1 R

R là ngôn ngữ lập trình mạnh mẽ cho phân tích thống kê và trực quan hóa dữ liệu, cung cấp nhiều gói hỗ trợ EDA như:

* ggplot2: Tạo các biểu đồ phức tạp và tùy chỉnh.
* dplyr: Xử lý dữ liệu hiệu quả.

### 2.2 MICROSOFT EXCEL

Excel cung cấp các công cụ cơ bản cho EDA, bao gồm:

* Hàm thống kê: Tính toán các chỉ số cơ bản.
* Biểu đồ: Tạo các biểu đồ trực quan.
* Pivot Table: Tóm tắt và phân tích dữ liệu.

**VÍ DỤ:**

### 2.3 TRỰC QUAN HÓA BIẾN ĐỘNG GIÁ XĂNG DẦU

A graph with blue bars

Description automatically generated with medium confidence

### 2.4 CHUẨN HÓA DỮ LIỆU

A screenshot of a computer

Description automatically generated

### 2.5 TÍNH TOÁN MA TRẬN

A screenshot of a computer

Description automatically generated

## **CHƯƠNG 3: TRỰC QUAN DỮ LIỆU**

### **1. PHÂN TÍCH THĂM DÒ DỮ LIỆU THAY ĐỔI XĂNG DẦU VIỆT NAM TRONG NĂM 2023 VÀ 2024 VỚI BIỂU ĐỒ PHÂN PHỐI CHUẨN**

### 1.1 GIỚI THIỆU

- Là phân phối xác suất liên tục cho một biến ngẫu nhiên. Biến ngẫu nhiên là biến mà giá trị của nó phụ thuộc vào kết quả của một sự kiện ngẫu nhiên. Có dạng hình chuông (bell curve), đối xướng qua giá trị trung bình; tức là trông giống như một ngọn đồi với đỉnh được xác định rõ, thì nó là phân phối chuẩn. Đỉnh của đường cong nằm ở mức trung bình và dữ liệu được phân phối đối xứng ở 2 bên của nó. Các giá trị trung bình, trung vị và mode bằng nhau hoặc nằm gần nhau.

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Sơ đồ

Mô tả được tạo tự độngẢnh có chứa biểu đồ, hàng, Sơ đồ, Phông chữ

Mô tả được tạo tự động

### 1.2 ĐỌC FILE DỮ LIỆU XĂNG DẦU

library(ggplot2)

# Đọc file CSV

data\_xangdau <- read.csv("/content/xangdau.csv")

print(data\_xangdau)

1.2 SỬ DỤNG THƯ VIỆN GGPLOT 2 VẼ BIỂU ĐỒ CHUẨN

#RON\_95\_V\_V1

ggplot(data\_xangdau[,c("DATE","RON\_95\_V\_V1")], aes( x = RON\_95\_V\_V1)) +

  geom\_histogram(aes(y=..density..), colour="black", fill="cyan",bins=30) +

  geom\_density(alpha=.2, fill="#FF6666") +

  labs(title="Biểu đồ phân phối chuẩn RON\_95\_V\_V1", x="Gia tri", y="Mat đo")

A graph of a diagram

Description automatically generated with medium confidence  
Biểu Đồ Phân Phối Chuẩn RON 95 V V1  
A graph of a diagram

Description automatically generated with medium confidence  
Biểu Đồ Phân Phối Chuẩn RON 95 III V1  
A graph with blue and pink lines

Description automatically generated  
Biểu Đồ Phân Phối Chuẩn E5 RON 92 II V1  
A graph with blue and pink lines

Description automatically generated  
A graph with blue and pink lines

Description automatically generatedBiểu Đồ Phân Phối Chuẩn Diezen 0,001 S-V V1  
  
Biểu Đồ Phân Phối Chuẩn Điêzen 0,005S II V1  
A graph with blue and pink lines

Description automatically generated  
A graph of a graph

Description automatically generated with medium confidenceBiểu Đồ Phân Phối Chuẩn Dầu Hỏa 2 K V1

Biểu Đồ Phân Phối Chuẩn Mazut N2B (3,5S) V1

A graph of a graph

Description automatically generated with medium confidence

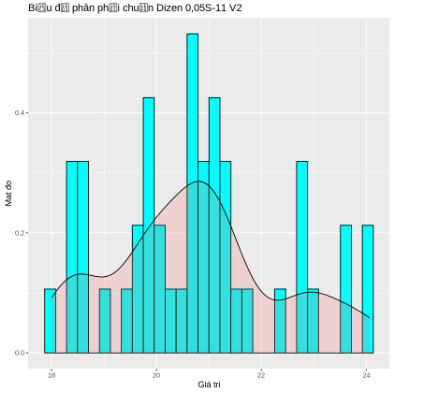
A graph with blue and pink lines

Description automatically generated  
Biểu Đồ Phân Phối Chuẩn Mazut 0,55 (RMG) V1

Biểu Đồ Phân Phối Chuẩn RON 95 V V2  
A graph with blue and pink lines

Description automatically generated  
A graph with blue and pink lines

Description automatically generatedBiểu Đồ Phân Phối Chuẩn RON 95 III V2  
  
Biểu Đồ Phân Phối Chuẩn E5 RON 92 II V 2

A graph with blue and pink bars

Description automatically generated  
Biểu Đồ Phân Phối Chuẩn Diezen S-V V2  
  
Biểu Đồ Phân Phối Chuẩn Diezen 0,005S II V2

A graph with blue and pink lines

Description automatically generatedA graph of a diagram

Description automatically generated with medium confidence  
Biểu Đồ Phân Phối Chuẩn Dầu Hỏa 2 – K V2

Biểu Đồ Phân Phối Chuẩn Mazut N2B (3,5S) V2

*A graph with blue and pink lines

Description automatically generated*  
Biểu Đồ Phân Phối Chuẩn Mazut 180cst-0,5S (RMG) V2

### 1.3 NHẬN XÉT

Các giá trị trung bình cho mỗi loại nhiên liệu và vùng không có sự thay đổi đáng kể giữa hai vùng. Độ lệch chuẩn cho thấy mức độ biến động của dữ liệu. Các loại nhiên liệu có độ biến động khá gần nhau. Phân phối của dữ liệu có vẻ không đều, với một số giá trị tập trung gần nhau trong một khoảng nhất định, nhưng cũng có sự biến động lớn ở các thời điểm khác nhau

### **2. PHÂN TÍCH THÀNH PHẦN CHÍNH -PCA**

### 2.1 GIỚI THIỆU

Trong bài viết này chúng em xin chia sẻ đến người đọc phương pháp phân tích thành phần chính (Principal Component Analysis - PCA), đây là một phương pháp được sử dụng thường xuyên khi các nhà phân tích thống kê phải đối mặt với những bộ số liệu với số chiều lớn (big data). Vậy làm thể nào để giảm thiểu chiều dữ liệu mà vẫn không mất đi thông tin và giữ lại được những thông tin cần thiết cho việc xây dựng các mô hình thì mời mọi người theo dõi bài viết dưới đây để hiểu rõ hơn phương pháp của PCA cũng như việcứng dụng R trong PCA đối với dữ liệu xăng dầu petrolimex .

Hình ảnh minh họa cho PCA:

Ảnh có chứa bóng rọi, động vật có vú, cái bóng

Mô tả được tạo tự động

Cùng là 1 chú lạc đà, tuy nhiên với các cách nhìn khác nhau (trục thông tin), chúng ta lại có những cách thu nhận thông tin khác nhau và cho ta những kết luận khác nhau.

**Hay nói cách khác, PCA là phương pháp chuyển hệ toạ độ (Hình mình họa)**

Hình 1: Thay đổi hệ toạ độ trong PCA

Ảnh có chứa hàng, biểu đồ, chữ viết tay, văn bản

Mô tả được tạo tự động

- PCA giả định rằng các hướng có phương sai lớn nhất là “quan trọng” nhất (tức là thành phần chính quan trọng nhất). Trong hình 2, trục PC1 là hướng chính đầu tiên mà các mẫu cho thấy sự biến thiên lớn nhất. Trục PC2 là hướng quan trọng thứ hai và nó trực giao với trục PC1.

Ảnh có chứa văn bản, biểu đồ, hàng, Sơ đồ

Mô tả được tạo tự động

Hình 2: Thay đổi hệ toạ độ trong PCA

**Eigenvalues / Variances**

* Eigenvalues: đo lường lượng biến thiên được giữ lại bởi mỗi thành phần chính. Eigenvalue lớn đối với các PC đầu tiên và nhỏ đối với các PC tiếp theo. Tức là, các PC đầu tiên tương ứng với các hướng có lượng biến thiên lớn nhất trong tập dữ liệu.
* Kiểm tra các giá trị Eigenvalues để xác định số lượng các thành phần chính được xem xét giữ lại. Các giá trị Eigenvalues và tỷ lệ phương sai được giữ lại bởi các thành phần chính (PC) có thể được trích xuất bằng cách sử dụng hàm get\_eigenvalue () [trong packages factoextra].

### 2.1 KHÁI NIỆM

- PCA là phương pháp biến đổi giúp giảm số lượng lớn các biến có tương quan với nhau thành tập ít các biến sao cho các biến mới tạo ra là tổ hợp tuyến tính của những biến cũ không có tương quan lần nhau. Ví dụ, chúng ta có 100 biến ban đầu có tương quan tuyến tính với nhau, khi đó chúng ta sử dụng phương pháp PCA xoay chiều không gian cũ thành chiều không gian mới mà ở đó chỉ còn 5 biến không có tương quan tuyến tính mà vẫn dữ được nhiều nhất lượng thông tin từ nhóm biến ban đầu.

### 2.2 ĐẶC TÍNH PCA

Một số đặc tính của PCA được kể đến như:

1. Giúp giảm số chiều dữ liệu - Giúp visualization khi dữ liệu có quá nhiều chiều thông tin.
2. Do dữ liệu ban đầu có số chiều lớn (nhiều biến) thì PCA giúp chúng ta xoay trục tọa độ xây một trục tọa độ mới đảm bảo độ biến thiên của dữ liệu và giữ lại được nhiều thông tin nhất mà không ảnh hưởng tới chất lượng của các mô hình dự báo. (Maximize the variability).
3. Do PCA giúp tạo 1 hệ trục tọa độ mới nên về mặt ý nghĩa toán học, PCA giúp chúng ta xây dựng những biến factor mới là tổ hợp tuyến tính của những biến ban đầu.
4. Trong không gian mới, có thẻ giúp chúng ta khám phá thêm những thông tin quý giá mới khi mà tại chiều thông tin cũ những thông tin quý giá này bị che mất (Điển hình cho ví dụ về chú lạc đà phía trên).

### 2.3 MÔ HÌNH PCA

Xét tập không gian (dữ liệu) k biến, k biến này được biểu qua j thành phần chính sao cho (j < k). Xét thành phần chính đầu tiên có dạng:

PC1=a1X1+a2X3+a4X5+...akXk

Thành phần chính đầu tiên chứa đựng hầu hết thông tin từ k biến ban đầu (được hình thành là 1 tổ hợp tuyến tính của các biến ban đầu) và lúc này tiếp tục xét thành phần chính thứ 2 được biểu diễn tuyến tính từ k biến ban đầu tuy nhiên thành phần chính thứ 2 phải không trực giao với thành phần chính ban đầu hay (thành phần chính thứ 2 không có mối tương quan tuyến tính với thành phần chính đầu tiên). Về lý thuyết chúng ta có thể xây dựng nhiều thành phần chính từ nhiều biến ban đầu. Tuy nhiên chúng ta cần tìm được trục không gian sao cho ít thành phần nhất mà có thể biểu diễn được hầu hết thông tin từ những biến ban đầu

### 2.4 THỰC HÀNH PHÂN TÍCH PCA TRONG R VỚI DỮ LIỆU XĂNG DẦU PETROLIMEX:

Cài các packages sử dụng trong PCA: install.packages(c("FactoMineR", "factoextra"))

**1.library**("FactoMineR")

**2.library**("factoextra")

**3.library**("corrr")

**4**. **library**("ggcorrplot")

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

**. Phân tích dữ liệu PCA Vùng I của dữ liệu xăng dầu PETROLIMEX**

- Chuẩn Hóa Dữ Liệu:

1.pca\_result <- prcomp(data\_1, scale = TRUE)

2.print(pca\_result)

A screenshot of a computer

Description automatically generatedKết quả :

-Tính Toán Ma Trận Tương Quan :

1.corr\_matrix\_1 <- cor(data\_1)

2.ggcorrplot(corr\_matrix\_1)

Kết quả :

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

Mô tả được tạo tự động

-Áp Dụng PCA:

**1.data\_1.pca <- princomp(corr\_matrix\_1)**

**2.summary(data\_1.pca)**

**Kết quả:**

**A screenshot of a computer

Description automatically generated**

+ Lấy dữ liệu :

A screenshot of a computer

Description automatically generated

+ Các biến:

“RON 95\_V1,E5\_RON 92\_1\_V1,

RCN\_95/111\_V1,Mazut\_180cst\_RMG\_V1,

Mazut\_N28\_V1,Diezen\_S\_I\_V1,Dauhga\_2\_KVI”

+Thành phần chính 1: Các giá trị dương của Comp.1 cho thấy sự tương quan dương giữa các biến RON\_95\_V\_V1, RON\_95\_111\_V1, E5\_RON\_92\_11\_V1, Dauhoa\_2\_KV1 và Mazut\_N2B\_V1. Các biến này có khả năng tương quan với nhau và đóng góp vào thành phần chính đầu tiên.

+Thành phần chính 2: Các giá trị dương của Comp.2 cho thấy sự tương quan dương giữa các biến RON\_95\_V\_V1, RON\_95\_111\_V1, E5\_RON\_92\_11\_V1 và Mazut\_180cst\_RMG\_V1. Tuy nhiên, các biến Diezen\_SV\_V1 và Diezen\_S\_ll\_V1 có tương quan âm với Comp.2. Thành phần chính thứ hai có thể liên quan đến sự phân tách giữa các loại nhiên liệu khác nhau. PCA giúp giảm chiều dữ liệu và tìm ra các thành phần quan trọng

+ Trực quan hóa các thành phần chính:

A graph with a line

Description automatically generated

+ Biplot của các thuộc tính:

A graph with lines and numbers

Description automatically generated

+ Đóng góp của mỗi biến :

A graph of a bar graph

Description automatically generated

+ Biplot kết hợp với cos2:

A screen shot of a graph

Description automatically generated

**Nhận xét**

**.**Biểu đồ này hiển thị các biến khác nhau như “RON 95\_V1”, “E5\_RON 92\_1\_V1”, “RCN\_95/111\_V1”, “Mazut\_180cst\_RMG\_V1”, “Mazut\_N28\_V1”, “Diezen\_S\_I\_V1” và các biến khác được vẽ trên hai thành phần chính - Dim1 (74,2%) và Dim2 (25,8%). Các trục được ghi nhãn là Dim1 và Dim2, cho thấy hai thành phần chính với tỷ lệ phương sai được giải thích trong ngoặc đơn.

**.**Có một thanh màu ở phía bên phải được ghi nhãn là ‘cos2’, có giá trị từ 0,020 đến 0,025, cho biết chất lượng biểu diễn của các biến trên bản đồ yếu tố. Biểu đồ sử dụng cả đường và nhãn văn bản để biểu thị vị trí của mỗi biến liên quan đến các thành phần chính này.

**Thành phần chính 1:** Các giá trị dương của Comp.1 cho thấy sự tương quan dương giữa các biến RON\_95\_V\_V1, RON\_95\_111\_V1, E5\_RON\_92\_11\_V1, Dauhoa\_2\_KV1 và Mazut\_N2B\_V1. Các biến này có khả năng tương quan với nhau và đóng góp vào thành phần chính đầu tiên.

**Thành phần chính 2:** Các giá trị dương của Comp.2 cho thấy sự tương quan dương giữa các biến RON\_95\_V\_V1, RON\_95\_111\_V1, E5\_RON\_92\_11\_V1 và Mazut\_180cst\_RMG\_V1. Tuy nhiên, các biến Diezen\_SV\_V1 và Diezen\_S\_ll\_V1 có tương quan âm với Comp.2. Thành phần chính thứ hai có thể liên quan đến sự phân tách giữa các loại nhiên liệu khác nhau. PCA giúp giảm chiều dữ liệu và tìm ra các thành phần quan trọng

**. Phân tích dữ liệu PCA Vùng II của dữ liệu xăng dầu PETROLIME:**

+ Chuẩn Hóa Dữ Liệu:

1.pca\_result <- prcomp(data\_2, scale = TRUE)

2.print(pca\_result)

Kết quả :

A screenshot of a computer

Description automatically generated

+ Tính Toán Ma Trận Tương Quan :

1.corr\_matrix\_2 <- cor(data\_2)

2.ggcorrplot(corr\_matrix\_2)

Ảnh có chứa văn bản, ảnh chụp màn hình, biểu đồ, Phông chữ

Mô tả được tạo tự độngKết quả :

+ Áp Dụng PCA:

**1.data\_2.pca <- princomp(corr\_matrix\_2)**

**2.summary(data\_2.pca)**

+ lấy dữ liệu

**A screenshot of a computer code

Description automatically generated**

Bảng cung cấp cho thấy sự quan trọng của các thành phần chính trong PCA.

.

-lấy dữ liệu :

A screenshot of a computer

Description automatically generated

-Trực quan hóa các thành phần chính:

A graph with a line

Description automatically generated

+ Biplot của các thuộc tính:

A screen shot of a graph

Description automatically generated

+ Đóng góp của mỗi biến:

A graph of a number of variable

Description automatically generated with medium confidence

+ Biplot kết hợp với Cos2:

A screen shot of a graph

Description automatically generated

**Nhận xét**

**các biến:**

“RON\_95\_V\_V2”: Biến này có giá trị âm trên PC1, PC3, và PC6, cho thấy nó có mối tương quan nghịch với các thành phần chính này. Nó cũng có giá trị dương trên PC4, PC5, PC7, và PC8, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“RON\_95\_111\_V2”: Biến này có giá trị âm trên PC1, PC3, và PC4, cho thấy nó có mối tương quan nghịch với các thành phần chính này. Nó cũng có giá trị dương trên PC5, PC6, PC7, và PC8, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“E5\_RON\_92\_11\_V2”: Biến này có giá trị âm trên PC1, PC3, PC4, và PC6, cho thấy nó có mối tương quan nghịch với các thành phần chính này. Nó cũng có giá trị dương trên PC5, PC7, và PC8, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“Diezen\_SV\_V2”: Biến này có giá trị âm trên PC1 và PC6, cho thấy nó có mối tương quan nghịch với các thành phần chính này. Nó cũng có giá trị dương trên PC2, PC4, PC5, và PC7, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“Diezen\_S\_ll\_V2”: Biến này có giá trị âm trên PC1, cho thấy nó có mối tương quan nghịch với thành phần chính này. Nó cũng có giá trị dương trên PC2, PC5, PC6, và PC8, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“Dauhoa\_2\_KV2”: Biến này có giá trị âm trên PC1 và PC6, cho thấy nó có mối tương quan nghịch với các thành phần chính này. Nó cũng có giá trị dương trên PC2, PC4, PC5, và PC7, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“Mazut\_N2B\_V2”: Biến này có giá trị âm trên PC1 và PC6, cho thấy nó có mối tương quan nghịch với các thành phần chính này. Nó cũng có giá trị dương trên PC2, PC4, PC5, và PC7, cho thấy nó có mối tương quan thuận với các thành phần chính này.

“Mazut\_180cst\_RMG\_V2”: Biến này có giá trị âm trên PC1 và lớn nhất trên PC2, cho thấy nó có mối tương quan nghịch với thành phần chính thứ nhất và mối tương quan thuận mạnh với thành phần chính thứ hai. Nó cũng có một mối tương quan thuận nhỏ với PC3 và PC8, cho thấy nó có một mối tương quan thuận nhẹ với các thành phần chính này.

**Bảng cung cấp cho thấy sự quan trọng của các thành phần chính trong PCA. Dưới đây là nhận xét chi tiết:**

Comp.1: Thành phần chính này có độ lệch chuẩn là 0.3801481, giải thích 72.57% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.1 chứa 72.57% thông tin của dữ liệu gốc.

Comp.2: Thành phần chính này có độ lệch chuẩn là 0.2076068, giải thích 21.64% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.2 chứa 21.64% thông tin của dữ liệu gốc.

Comp.3: Thành phần chính này có độ lệch chuẩn là 0.10279984, giải thích 5.31% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.3 chứa 5.31% thông tin của dữ liệu gốc.

Comp.4: Thành phần chính này có độ lệch chuẩn là 0.029492249, giải thích 0.44% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.4 chứa 0.44% thông tin của dữ liệu gốc.

Comp.5: Thành phần chính này có độ lệch chuẩn là 0.0078451605, giải thích 0.03% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.5 chứa 0.03% thông tin của dữ liệu gốc.

Comp.6: Thành phần chính này có độ lệch chuẩn là 4.099863e-03, giải thích 0.0084% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.6 chứa 0.0084% thông tin của dữ liệu gốc.

Comp.7: Thành phần chính này có độ lệch chuẩn là 2.750177e-03, giải thích 0.0038% tổng biến động của dữ liệu. Điều này có nghĩa là Comp.7 chứa 0.0038% thông tin của dữ liệu gốc.

Comp.8: Thành phần chính này có độ lệch chuẩn rất nhỏ (3.232333e-09), giải thích một phần rất nhỏ (5.246638e-17%) tổng biến động của dữ liệu. Điều này có nghĩa là Comp.8 chứa một lượng thông tin rất nhỏ của dữ liệu gốc.

Tổng cộng, tất cả các thành phần chính đã giải thích được 100% tổng biến động của dữ liệu, cho thấy rằng chúng đã giữ được toàn bộ thông tin của dữ liệu gốc, mặc dù số lượng biến đã được giảm xuống.

biểu đồ cung cấp, Dim1 và Dim2 chính là hai thành phần chính đầu tiên. Dim1 giải thích 72.6% tổng biến động của dữ liệu, trong khi Dim2 giải thích 22.6%. Các điểm dữ liệu trên biểu đồ thể hiện mối tương quan giữa các biến gốc và hai thành phần chính này. Mỗi biến gốc được biểu diễn bởi một vector, và độ dài của vector cho biết mức độ quan trọng của biến đó trong việc giải thích biến động của dữ liệu. Màu sắc của vector cho biết giá trị cos2, thể hiện mức độ mà biến gốc được biểu diễn bởi thành phần chính. Một giá trị cos2 cao cho thấy biến đó được biểu diễn tốt bởi thành phần chính.

biểu đồ cột với tiêu đề “Cos2 của các biến đối với Dim1-2”. Biểu đồ này thể hiện giá trị Cos2, chất lượng của biểu diễn, của các biến trên trục x. Các biến bao gồm “FROB.SY > V”, “Malt ALE.SY > V”, “GREEN S > V”, “Diversey > V”, “FROB.NA > V”, “ESPOIR.SY.LT > V”, và “Malt.NA > V”.

Mỗi biến trên trục x tương ứng với một cột màu xanh biểu diễn giá trị Cos2 của nó. Các cột không đồng đều về chiều cao, cho thấy mức độ chất lượng biểu diễn khác nhau cho mỗi biến.

từ biểu đồ này, chúng ta có thể thấy rằng một số biến được biểu diễn tốt hơn bởi Dim1 và Dim2 so với các biến khác. Tuy nhiên, tất cả các giá trị Cos2 đều khá thấp (dưới 0.03), cho thấy không có biến nào được biểu diễn rất tốt bởi Dim1 và Dim2. Điều này có thể cho thấy rằng có thể cần thêm thành phần chính để biểu diễn đầy đủ các biến gốc.

“RON 95 V2”: Điểm dữ liệu này nằm gần trục Dim1, cho thấy nó có mối tương quan mạnh mẽ với Dim1. Tuy nhiên, nó cũng có một khoảng cách nhất định từ trục Dim2, cho thấy nó cũng có một mối tương quan với Dim2, mặc dù không mạnh mẽ như Dim1.

“E5 RON 92\_1 V2”: Điểm dữ liệu này nằm gần trục Dim2, cho thấy nó có mối tương quan mạnh mẽ với Dim2. Tương tự như “RON 95 V2”, nó cũng có một khoảng cách nhất định từ trục Dim1, cho thấy nó cũng có một mối tương quan với Dim1, mặc dù không mạnh mẽ như Dim2.

“Mazut 180CST RMG\_V2”: Điểm dữ liệu này nằm gần gốc của hệ trục tọa độ, cho thấy nó có mối tương quan vừa phải với cả Dim1 và Dim2.

## **KẾT LUẬN**

Cùng với sự hướng dẫn và hỗ trợ nhiệt tình của thầy giáo và tham khảo trên mạng ,một kỳ học là khoảng thời gian quý báu để nhóm em có những trải nghiệm mới lạ và thú vị , rèn luyện , áp dụng được những kiến thức thầy cô chỉ bảo , để rồi có cơ hội trau dồi kiến thức và gặt hái thêm kinh nghiệm , tạo nền tảng , giúp nhóm em ngày càng hoàn thiện hơn . Mặc dù trong qua trình làm đồ án vẫn còn nhiều sự bỡ ngỡ và chưa đủ kiến thức và kĩ năng , nhưng nhóm em cũng đã cố gắng hoàn thành một cách chỉnh chu phù hợp với yêu cầu thầy đưa ra

**Tổng kết :**

- Tìm hiểu được cách phân tích dữ liệu của cá nhân hoặc một tập thể

- Tìm hiểu và áp dụng được kiến thức để sử dụng biểu đồ phân phối chuẩn dành cho việc phân tích dữ liệu   
- Tìm hiểu và áp dụng được cách phân tích dữ liệu PCA dành cho những số liệu thống kê lớn thành các dữ liệu nhỏ giúp người làm dễ dàng hơn

|  |
| --- |
|  |

## **TÀI LIỆU THAM KHẢO**

|  |
| --- |
| [1]. Phan Duy Hùng, Đỗ Thái Giang (2022), Ngôn ngữ lập trình R, Nhà xuất bản ĐH |
| Quốc gia Hà Nội. |
| 21. 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]. Emmanuel Paradis (2005), R for Beginners. |
| Link: https://cran.r-project.org/doc/contrib/Paradis-rdebuts\_en.pdf |
| [6]. J. H. Maindonald (2008), Using R for Data Analysis and Graphics Introduction, Code and Commentary. |
| Link: https://cran.r-project.org/doc/contrib/usingR.pdf |
| [7]. John Verzani (2005), Used R for Introductory Statistics. |
| Link: https://cbb.sjtu.edu.cn/~mywu/bi217/usingR.pdf |
| [8]. Jonathan Baron và Yuelin Li. Web (2011), Notes on the use of R for psychology experiments and questionnaires. |
| Link:https://www.sas.upenn.edu/~baron/from\_cattell/rpsych/rpsych.html |
| [9]. Julian J. Faraway (2002), Practical Regression and Anova using R. |
| Link: |
| https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf |
| [10]. W. N. Venables, D. M. Smith and the R Core Team (2022), An Introduction to R. |
| Link: |
| http://cran.r-project.org/doc/manuals/R-intro.pdf. |
| [11]. Alexey Shipunov (2019), Visual Statistics Use R! |
| Link: https://cran.r-project.org/doc/contrib/Shipunov-visual\_statistics.pdf |
| [12]. Eric Pimpler (2017), Data Visualization and Exploration with R. |
| Link: https://www.passeidireto.com/arquivo/117834003/data-visualization-and- |
| exploration-with-r-a-practical-guide-to-using-r-r-studio- |
| [13]. Web: https://rmarkdown.rstudio.com/authoring\_quick\_tour.html. |
| [14]. Web: https://r-graph-gallery.com/ |
| [15]. Web: http://thongke.cesti.gov.vn/dich-vu-thong-ke/tai-lieu-phan-tich-thong-ke |