BỘ CÔNG THƯƠNG
TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO ĐỒ ÁN
"SỬ DỤNG NGÔN NGỮ R ĐỂ THĂM DÒ VÀ TRỰC QUAN DỮ LIỆU VỀ GIÁ CHỨNG KHOÁN CÁC MÃ LIÊN QUAN CỦA CÔNG TY CPTM VÀ KHAI THÁC KHOÁNG SẢN DƯƠNG HIẾU"

Stt	Họ và tên sinh viên	Mã sinh viên
1	Trần Đức Trung	22174600059
2	Phạm Văn Thắng	22174600100
3	Đậu Thị Thảo	22174600003
4	Trần Thị Thảo Vân	22174600019
5	Khuất Thanh Phương	22174600005

Hà Nội – 2024

BỘ CÔNG THƯƠNG

TRƯỜNG ĐẠI HỌC KINH TẾ KỸ THUẬT CÔNG NGHIỆP

KHOA KHOA HỌC ỨNG DỤNG

BÁO CÁO ĐỒ ÁN

"SỬ DỤNG NGÔN NGỮ R ĐỂ THĂM DÒ VÀ TRỰC QUAN DỮ LIỆU VỀ GIÁ CHỨNG KHOÁN CÁC MÃ LIÊN QUAN CỦA CÔNG TY CPTM VÀ KHAI THÁC KHOÁNG SẢN DƯƠNG HIẾU"

Stt	Họ và tên sinh viên	Mã sinh viên
1	Trần Đức Trung	22174600059
2	Phạm Văn Thắng	22174600100
3	Đậu Thị Thảo	22174600003
4	Trần Thị Thảo Vân	22174600019
5	Khuất Thanh Phương	22174600005

Hà Nội – 2024

DANH MỤC HÌNH ẢNH BẢNG BIỂU

LỜI CẢM ƠN

Trước hết, nhóm chúng em xin bày tỏ tình cảm và lòng biết ơn tới thầy ThS. Trần Chí Lê- người đã từng bước hướng dẫn, giúp đỡ chúng em trong suốt quá trình thực hiện bài báo cáo cuối kỳ.

Em xin chân thành cảm ơn thầy đã dẫn dắt, dạy dỗ chúng em về cả kiến thức chuyên môn và tinh thần học tập để có những kiến thức để thực hiện đồ án cuối kỳ của mình.

Tuy có nhiều cố gắng trong quá trình học tập cũng như trong quá trình làm đồ án không thể tránh khỏi những thiếu sót, chúng em rất mong có được sự góp ý quý báu của thầy cũng như tất cả các anh chị để kết quả của chúng em được hoàn thiện hơn.

Một lần nữa chúng em xin chân thành cảm ơn.

Hà Nội, ngày 19 tháng 05 năm 2024

Nhóm thực hiện đồ án - Nhóm 2.1

LỜI MỞ ĐẦU

Tính đến cuối tháng 2/2024, Việt Nam có gần 7,5 triệu tài khoản chứng khoán cá nhân, tương đương với khoảng 7.5% dân số. Như vậy phân tích chứng khoán doanh nghiệp là một bước quan trọng và cần thiết trong quá trình đầu tư vào thị trường này. Điều này giúp đánh giá giá trị thực của cổ phiếu, hiểu rõ về tình hình tài chính và hoạt động của doanh nghiệp, đồng thời cũng dự báo triển vọng tương lai dựa trên các yếu tố như chiến lược kinh doanh hay mô hình, hơn hết phân tích chứng khoán giúp nhà đầu tư quản lý rủi ro, lựa chọn danh mục đầu tư phù hợp. Tóm lại, phân tích chứng khoán doanh nghiệp là một công cụ không thể thiếu giúp nhà đầu tư đưa ra các quyết định sáng suốt và hiệu quả, tối ưu hóa lợi nhuận quản lý rủi ro trong môi trường đầy biến động.

Báo cáo phân tích này sử dụng ngôn ngữ R để thăm dò và trực quan dữ liệu về mã chứng khoán công ty cổ phần Thương mại và Khai thác khoáng sản Dương Hiếu (DHM) . Ngoài ra báo cáo cũng so sánh, đánh giá các dữ liệu và phân tích mối liên hệ giữa các giá trị dữ liệu. Từ đó đưa ra nguyên tố, nhóm các thành phần gây ảnh hưởng lớn đến các thành phần còn lại của chứng khoán công ty nhằm nâng cao chất lượng đầu tư và giúp công ty hiểu rõ về môi trường kinh doanh, tình hình tài chính của mình và đưa ra quyết định chiến lược thông minh để tối ưu hóa hiệu suất kinh doanh.

Về kết cấu, báo cáo gồm 4 phần như sau:

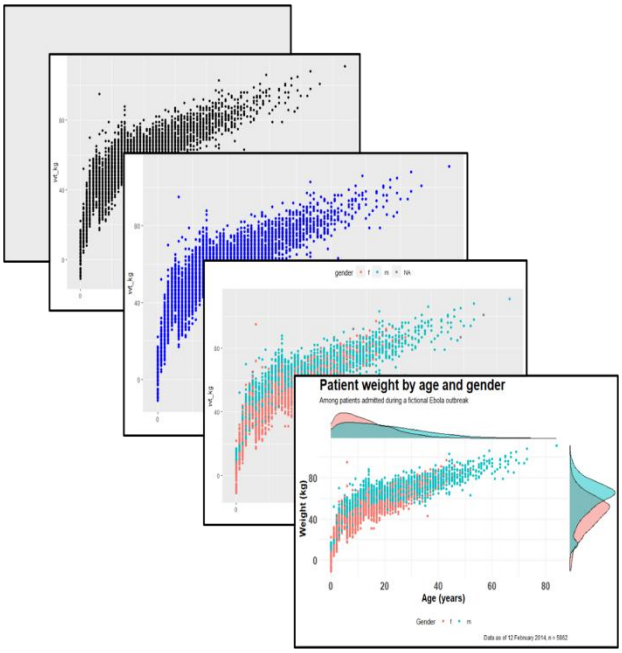
- Chương 1: R và các kỹ thuật thăm dò, trực quan dữ liệu cơ bản.
- Chương 2: Phân tích, thăm dò dữ liệu
- Chương 3: Trực quan hóa dữ liệu
- Chương 4: Tổng kết

CHƯƠNG 1: R VÀ CÁC KỸ THUẬT THĂM DÒ, TRỰC QUAN DỮ LIỆU CƠ BẢN

Phần 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 các 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)

1.1. THĂM DÒ DỮ LIỆU BẰNG BẢNG BIỂU

1.1.1. BIỂU ĐỒ VỚI BACKAGE GGPLOT



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.

```
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         a4      2.8  1999     6 manual(m5) f        18    26 p     compa…
#> # ℹ 228 more rows
```

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:

```
ggplot(data = my_data)+                  # sử dụng dữ liệu "my_data"
  geom_yyy(                             # thêm một lớp các hàm-hình biểu đồ
    mapping = aes(x = col1, y = col2),  # gán dữ liệu tới các trục
    color = "red")+                     # thêm một số đặc điểm khác (như màu sắc)
  labs()+                               # thêm tiêu đề, nhãn, bảng số,..
  theme()                               # điều chỉnh cỡ chữ, màu sắc, phông chữ
```

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.

- 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.

- 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:

- 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, hoặc 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 theme()	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.

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 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() (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.

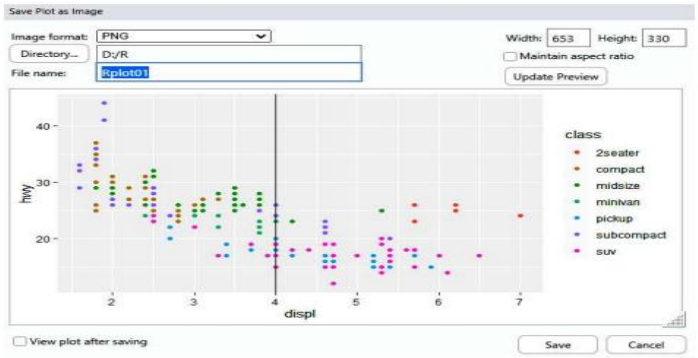
o Với hàm ggsave(), có thể được tiến hành theo hai cách:

▪ Chỉ định tên của đối tượng biểu đồ, sau đó là đường dẫn tệp và tên có phần mở rộng. Ví dɔ: 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”, hohc “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 1.17: Xuất bản biểu đồ qua lệnh Import trong Rstudio.

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.

1.1.2. KHÁI QUÁT THĂM DÒ DỮ LIỆU

Thăm dò dữ liệu EDA (Exploratory Data Analysis) là quá trình mô tả dữ liệu bằng các kỹ thuật thống kê và trực quan hóa nhằm tập trung vào khía cạnh quan trọng của dữ liệu để tiếp tục phân tích. Điều này bao gồm cả việc kiểm tra tập dữ liệu từ nhiều góc độ, mô tả và tóm tắt nó mà không đưa ra bất cứ giả định nào khác về nội dung của nó. EDA là một bước rất quan trọng cần phải thực hiện trước khi đi sâu hoặc học máy. Có bốn loại phân tích thăm dò dữ liệu là: đơn biến phi đồ họa; Đa biến phi đồ họa; Đồ họa đơn biến và Đồ họa đa biến.

1.2. TÓM TẮT KẾT QUẢ THEO SUY DIỄN

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

1.2.1. KHÁI QUÁT THỐNG KÊ MÔ TẢ

Cho một biến số $x_1, x_2, x_3, ..., x_n$ chúng ta có thể tính toán một số chỉ số thống kê mô tả như sau:

Lý thuyết	Hàm R
Số trung bình: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	`mean(x)`
Phương sai: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$	`var(x)`
Độ lệch chuẩn: $s = \sqrt{s^2}$	`sd(x)`
Trị số thấp nhất	`min(x)`
Trị số cao nhất	`max(x)`
Toàn cự (range)	`range(x)`

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 $x_1, x_2, ..., x_n$ được chọn từ tổng thể có phân phối chuẩn (hoặc xấp xỉ chuẩn tức phân phối có dạng đối xứng) với kỳ vọng a và phương sai σ^2.

Giả thuyết H_0: $a = a_0$; Đối thuyết H_1: $\begin{cases} a \neq a_0 \\ a < a_0 \\ a > a_0 \end{cases}$ (Một trong 3 trường hợp)

Tính thống kê kiểm định: $t = \frac{\bar{x} - a_0}{s}.\sqrt{n}$.

Miền bác bỏ:

- Với H_1: $a \neq a_0$, bác bỏ H_0 nếu $t < -t_{1-\alpha/2}^{n-1}$ hoặc $t > t_{1-\alpha/2}^{n-1}$.
- Với H_1: $a < a_0$, bác bỏ H_0 nếu $t < -t_{1-\alpha}^{n-1}$.
- Với H_1: $a > a_0$, bác bỏ H_0 nếu $t > t_{1-\alpha}^{n-1}$.

Trong R, để tìm phân vị $t_{1-\alpha/2}^{n-1}$ sử dụng hàm `qt(1-alpha/2,n-1)`.

Trong kết quả do R xuất ra, ta xác định có bác bỏ H_0 hay không thông qua P- giá trị.

Quy tắc: Khi P- giá trị bé hơn α thì bác bỏ H_0.

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

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

```
t.test(x,alternative= c("two.sided", "less", "greater"),mu = mu_0, conf.level = 0.95)
# Trong đó:
```

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

o x: véc tơ dữ liệu.

o alternative: xác đlnh 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: $f = \frac{m}{n}$. Giả thuyết cỡ mẫu khảo sát n phải tương đối lớn.

Giả thuyết: $H_0: p = p_0$; Đối thuyết: $H_1: \begin{cases} p \neq p_0 \\ p < p_0 \\ p > p_0 \end{cases}$ (Một trong 3 trường hợp).

Tính thống kê kiểm định: $u = \frac{f - p_0}{\sqrt{p_0.(1-p_0)}}.\sqrt{n}$.

Miền bác bỏ:

- Với $H_1: p \neq p_0$ bác bỏ H_0 nếu $u < -z_{1-\alpha/2}$ hoặc $u > z_{1-\alpha/2}$.
- Với $H_1: p < p_0$ bác bỏ H_0 nếu $u < -z_{1-\alpha}$.
- Với $H_1: p > p_0$ bác bỏ H_0 nếu $u > z_{1-\alpha}$.

Để tìm $z_{1-\alpha/2}$, sử dụng hàm qnorm(1-alpha/2).

Sử dụng hàm prop.test để kiểm định:

```
prop.test(m, n, p = p0, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
# trong đó:
```

Sử dụng hàm prop.test để kiểm định:

```
prop.test(m, n, p = p0, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
```

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:

```
> t.test(T~gioitinh, var.equal = TRUE)
Kết quả hiển thị:
        Two Sample t-test
data:  T by gioitinh
t = 0.51659, df = 28, p-value = 0.6095
alternative hypothesis: true difference in means between group Nam and group Nu is not
equal to 0
95 percent confidence interval:
 -0.5139801  0.8606468
sample estimates:
mean in group Nam  mean in group Nu
         7.306667          7.133333
```

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 tỉ lệ**

Cho hai mẫu với số đối tượng n_1 và n_2, gọi số phần tử thỏa mãn tính chất A trong mẫu 1 là m_1, trong mẫu 2 là m_2. Do đó, chúng ta có thể tính được tỉ lệ tương ứng trong hai mẫu là p_1, p_2. Lí thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu $d = p_1 - p_2$ tuân theo luật phân phối chuẩn với số trung bình 0 và phương sai bằng:

$V_d = (\frac{1}{n_1} + \frac{1}{n_2})\mathrm{p}(1-\mathrm{p})$. Trong đó: $\mathrm{p} = \frac{m_1 + m_2}{n_1 + n_2}$; $z = d / V_d$ tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1.

Kết quả phân tích trên cho thấy tỷ lệ gãy xương trong nhóm 1 là 0.07 và nhóm 2 là 0.18. Phân tích này cũng cho thấy xác suất 95% rằng độ khác biệt giữa hai nhóm có thể là từ 0.01 đến 0.20 (tức từ 1 đến 20%). Với giá trị p = 0.027, chúng ta có thể nói rằng tỉ lệ gãy xương trong nhóm A thực sự thấp hơn nhóm B.

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)

Chúng tôi đưa ra bài toán 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:

```
> var.test(T~gioitinh)
Kết quả hiển thị:
F test to compare two variances
data:  T by gioitinh
F = 0.45106, num df = 14, denom df = 14, p-value = 0.1485
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1514355 1.3435331
sample estimates:
ratio of variances
         0.4510638
```

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:

```
shapiro.test(x)
trong đó: x: là dữ liệu mẫu
```

Hình 1.8: Thủ tục kiểm định

1.2.3. THỐNG KÊ SUY DIỄN TRONG CÁC BÀI PHÂN TÍCH LIÊN 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 "markettimng.csv"có thể tham khảo từ https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu

```
library(readr)
marketing <- read_csv "marketing.csv"
head(marketing)
# kết quả hiển thi
```

Hình 1.9: Đọc file marketing.csv

1	youtube	facebook	newspaper	sales
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	201.	142.	145.	943.
2	2	156.	130.	62.7	856.
3	3	124.	188.	140.	965.
4	4	158.	187.	144.	1017.
5	5	158.	222.	116.	1115.
6	6	132.	182.	120.	932.
7	7	121.	214.	144.	1022.
8	8	108.	82.6	126.	650.
9	9	190.	173.	104.	1001.
10	10	117.	115.	133.	713.
# ... with 190 more rows					

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

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

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:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Trong đó, $\bar{x}$ và $\bar{y}$ là giá trị trung bình của biến số x và y.

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

```
cor(data, method = "pearson")
```

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:

```
cor(data, method = " spearman ")
```

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:

```
cor(data, method = " kendall ")
```

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

- **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:

```
cor.test(nhân tố 1, nhân tố 2, method = c("pearson","spearman", "kendall"))
```

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.

Method được lựa chọn một trong ba phương pháp tương ứng.

CHƯƠNG 2: PHÂN TÍCH, THĂM DÒ DỮ LIỆU

2.1. BÀI TOÁN PHÂN TÍCH ĐỀ XUẤT

Trình bày phần mở đầu đã nêu rõ về sự tăng mạnh dân số Việt Nam tham gia đầu tư chứng khoán- thị trường có nhiều biến động.

Báo cáo phân tích này sử dụng ngôn ngữ R để thăm dò và trực quan dữ liệu về mã chứng khoán công ty cổ phần Thương mại và Khai thác khoáng sản Dương Hiếu (DHM) . Ngoài ra báo cáo cũng so sánh, đánh giá các dữ liệu và phân tích mối liên hệ giữa các giá trị dữ liệu. Từ đó đưa ra nguyên tố, nhóm các thành phần gây ảnh hưởng lớn đến các thành phần còn lại của chứng khoán công ty nhằm nâng cao chất lượng đầu tư và giúp công ty hiểu rõ về môi trường kinh doanh, tình hình tài chính của mình và đưa ra quyết định chiến lược thông minh để tối ưu hóa hiệu suất kinh doanh.

2.1.1. DỮ LIỆU

2.1.1.1. Thu thập dữ liệu

Sử dụng nguồn dữ liệu phổ biến của web CafeF https://s.cafef.vn/hose/dhm-cong-ty-co-phan-thuong-mai-va-khai-thac-khoang-san.chn

Bằng việc tìm kiếm keywords kết hợp ngôn ngữ lập trình cùng keywords tiếng anh. Dữ liệu này tập trung vào giá chứng khoán và các mã liên quan của công ty cptm Dương Hiếu và không bao gồm các dữ liệu từ doanh nghiệp khác. Vì thế các phân tích trong bài báo cáo này sẽ dựa trên thông tin hiện có và có một số giới hạn về tính chính xác đầy đủ của dữ liệu..

Các phương pháp thống kê mô tả và trực quan hóa dữ liệu để phân tích dữ liệu từ các mã chứng khoán xuất hiện trong dữ liệu được thu thập. Đồng thời các công cụ và ggplot2 trong R cũng sẽ được sử dụng để thực hiện các phân tichs và tạo biểu đồ.

2.1.1.2. Biến dữ liệu

Để thuận tiện cho việc phân tích, chúng tôi đặt tên các biến dữ liệu bằng các ký tự như sau.

Ký tự	Biến dữ liệu	Ký tự	Biến dữ liệu
A	Năm	I	Phải thu ngắn hạn của khách hàng
B	Tài sản ngắn hạn	J	Trả trước cho người bán ngắn hạn
C	Tiền và các tài sản tương đương	M	Phải thu ngắn hạn khác
D	Tiền	O	Hàng tồn kho
E	Đầu tư tài chính ngắn hạn	Q	Tài sản ngắn hạn khác
F	Chứng khoán kinh doanh	BE	Lợi nhuận phân phối sau thuế
H	Khoản thu ngắn hạn của	BI	Tổng cộng nguồn vốn
R	Chi phí trả trước ngắn hạn	AG	Tài sản cố định vô hình
S	Thuế GTGT được khấu trừ	AK	Tài sản dở dang dài hạn
U	Tài sản dài hạn	AR	Tài sản và chi phí
AB	Tài sản cố định	AV	Tổng cộng tài sản
AC	Tài sản cố định hữu hình	AW	Nợ pahir trả

AX	Nợ ngắn hạn	BD	Quỹ đầu tư phát triển
BB	Vốn chủ sở hữu		

Tài sản ngắn hạn: là các tài sản dự kiến được chuyển đổi thành tiền mặt hoặc sử dụng trong vòng một năm hoặc chu kỳ hoạt động bình thường của công ty. Các tài sản này có tính thanh khoản cao hơn và thường được sử dụng để tài trợ cho các hoạt động ngắn hạn. Các loại tài sản ngắn hạn phổ biến bao gồm:

- Tiền mặt và các khoản tương đương tiền: Gồm tiền mặt, séc, tài khoản ngân hàng và các khoản tương đương tiền khác. Các khoản phải thu: Khoản tiền khách hàng nợ công ty. Hàng tồn kho: Bao gồm nguyên vật liệu, hàng hóa đang trong quá trình sản xuất và thành phẩm. Chứng khoán đầu tư ngắn hạn: Đầu tư có tính thanh khoản cao có thể được chuyển đổi thành tiền mặt nhanh chóng. Tài sản dài hạn: Tài sản dài hạn là các tài sản mà công ty dự kiến giữ lâu hơn một năm hoặc chu kỳ hoạt động bình thường. Các tài sản này có tính thanh khoản thấp hơn và thường được sử dụng cho các hoạt động dài hạn. Các loại tài sản dài hạn phổ biến bao gồm:

- Tài sản cố định: Gồm đất đai, nhà cửa, máy móc, thiết bị và các công cụ khác phục vụ hoạt động kinh doanh dài hạn. Tài sản vô hình: Bao gồm bằng sáng chế, thương hiệu, giấy phép và các tài sản vô hình khác. Đầu tư dài hạn: Khoản đầu tư mà công ty dự kiến giữ lâu dài hoặc không dự định chuyển đổi thành tiền mặt nhanh chóng. Mối tương quan giữa tài sản ngắn hạn và dài hạn: Mối tương quan giữa tài sản ngắn hạn và dài hạn phản ánh chiến lược tài chính và kinh doanh của công ty:

Công ty có nhiều tài sản ngắn hạn: Cho thấy công ty có tính thanh khoản cao, dễ dàng đáp ứng nhu cầu tài chính trong ngắn hạn. Tuy nhiên, điều này có thể chỉ ra rằng công ty chưa tận dụng tối đa nguồn lực để đầu tư vào tài sản dài hạn. Công ty có nhiều tài sản dài hạn: Cho thấy công ty đang đầu tư mạnh vào tài sản cố định và phát triển dài hạn, nhưng có thể gặp khó khăn về thanh khoản

trong ngắn hạn. Cân bằng giữa tài sản ngắn hạn và dài hạn: Phản ánh sự cân bằng giữa tính thanh khoản và đầu tư dài hạn, cho thấy công ty đang cố gắng quản lý rủi ro tài chính và đầu tư. Nhìn chung, việc xem xét cấu trúc tài sản ngắn hạn và dài hạn có thể giúp hiểu rõ hơn về chiến lược kinh doanh của công ty, rủi ro tài chính và khả năng sinh lợi trong ngắn hạn và dài hạn.

Một số các thuật ngữ khác chúng tôi sẽ tiến hành chú thích trong quá trình phân tích.

2.1.2. XỬ LÝ DỮ LIỆU

Sử dụng hàm ‘read_excel’ để nhập dữ liệu vào môi trường phân tích:

```
install.packages("dplyr")
library(dplyr)
library(readxl)
library(ggplot2)
library(tidyr)
data <- read_excel("/content/doan1.xlsx")
do_an1 <- data
do_an1

```

Kết quả hiển thị:

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

A tibble: 10 × 29

Năm	B	C	D	E	F	H	I	J	M	⋯	AR	AS	AV	AW	AX	BA	BB	BD	BE	BI
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	⋯	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Quý 42023	691.317	12.110	12.110	44.532	41.600	545.704	469.068	16.013	64.271	⋯	393	393	899.277	550.188	549.188	349.089	349.089	31.018	4.253	899.277
Quý 32023	338.194	14.836	14.836	47.431	44.738	186.464	90.059	63.598	36.456	⋯	394	394	546.949	198.980	197.980	347.969	347.969	31.018	3.133	546.949
Quý 22023	359.224	13.354	13.354	47.671	42.197	214.530	87.146	64.377	48.655	⋯	361	361	568.472	221.308	220.308	347.163	347.163	31.018	2.327	568.472
Quý 12023	314.104	25.142	17.942	50.319	52.046	160.520	116.717	16.500	38.451	⋯	387	387	541.646	195.062	195.062	346.585	346.585	31.018	1.749	541.646
Quý 42022	363.190	3.311	3.311	45.364	54.511	224.308	151.929	55.651	35.134	⋯	472	472	569.451	218.659	218.659	350.793	350.793	31.018	5.957	569.451
Quý 32022	323.433	6.075	5.575	49.978	59.125	166.271	91.927	82.389	10.361	⋯	416	416	528.383	180.787	179.787	347.596	347.596	31.018	2.760	528.383
Quý 22022	471.726	3.064	3.064	52.858	58.747	331.570	305.483	9.714	34.779	⋯	438	438	677.198	330.419	330.419	346.778	346.778	31.018	1.942	677.198
Quý 12022	330.408	32.833	21.021	108.308	64.336	110.594	61.302	66.430	1.263	⋯	427	427	536.391	190.995	190.995	345.396	345.396	31.018	560.000	536.391
Quý 42021	270.486	1.564	1.564	49.867	49.867	85.970	94.358	8.498	1.514	⋯	465	465	582.955	271.290	271.290	311.665	311.665	31.219	-33.371	582.955
Quý 32021	364.019	42.133	2.133	NA	NA	120.366	68.384	37.736	32.646	⋯	498	498	696.152	417.834	364.834	278.318	278.318	31.219	-66.719	696.152

Bảng 2.1: Dữ liệu chứng khoán công ty CPTM từ năm 2021-2023

Tiến hành xử lý dữ liệu, thay thế các giá trị NA(các giá trị không tồn tại, hoặc không có giá trị) bằng giá trị 0. Chương trình như sau:

```r
View(do_an1)
#thay thế giá trị na = 0
is.na(do_an1)
do_an1$E[is.na(do_an1$E)] <- 0
do_an1$F[is.na(do_an1$F)] <- 0
do_an1$U[is.na(do_an1$U)] <- 0
do_an1$AG[is.na(do_an1$AG)] <- 0
View(do_an1)
do_an1
```

Khi hiển thị dữ liệu trong dataframe dòng 'is.na(do_an1)' sẽ kiểm tra xem có giá trị NA (rỗng) nào trong data frame do_an1 hay không, đồng thời các dòng sau sử dụng kỹ thuật indexing để chọn ra các phần tử có giá trị NA trong từng cột E, F, U, AG, và sau đó thay thế chúng bằng 0.

Cách xử lý trên cho phép kiểm tra các giá trị NA đã được thay thế thành 0 một cách chính xác hay không, giúp đảm bảo tính chính xác và toàn vẹn của dữ liệu.

2.2 MỐI LIÊN HỆ GIỮA CÁC BIẾN

2.2.1. Ma trận tương quan

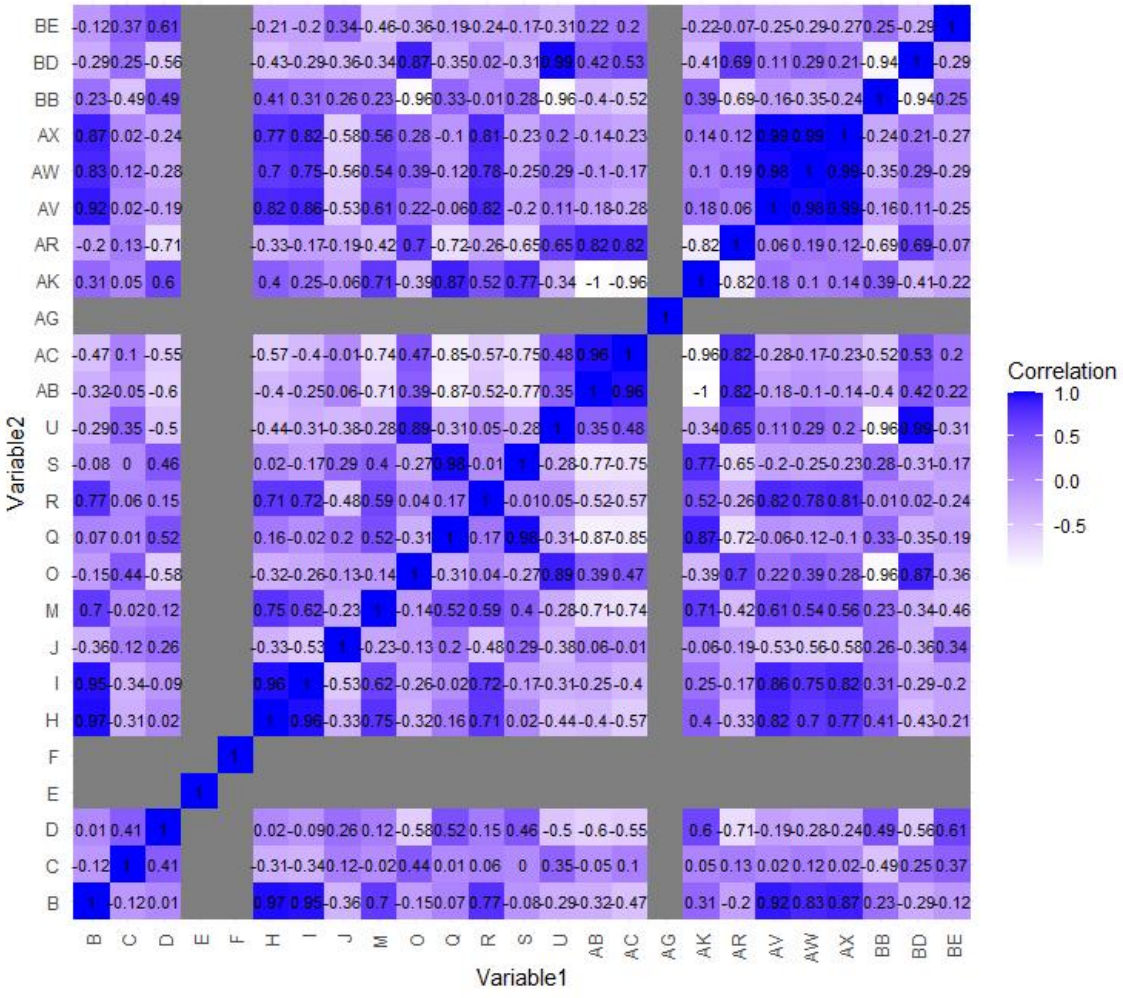
Khám phá mối quan hệ giữa các biến dữ liệu, chúng ta sẽ bắt đầu bằng việc tính ma trận tương quan. Ma trận tương quan cung cấp cái nhìn toàn diện về mức độ và hướng của mối quan hệ tuyến tính giữa các biến. Để tính ma trận tương quan trong R, ta dùng cú pháp sau:

```r
cor_matrix <- cor(data)
```

2.2.2 Ma trận tương quan và biểu đồ về các biến chứng khoán

Sử dụng biểu đồ heatmap để trực quan hóa ma trận tương quan đồng thời vẽ biểu đồ giá của từng mã chứng khoán để thấy rõ sự biến động giá theo thời gian.

```r
# Xoá cột 'Năm' để chỉ giữ lại dữ liệu số
data_numeric <- data[-1]
# Tính ma trận tương quan
correlation_matrix <- cor(data_numeric)
# 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))
```



Hình 2.1 Ma trận tương quan giữa các biến

Nhận xét:

- correlation_matrix <- cor(data_numeric): Ma trận tương quan này sẽ cho ta biết mối tương quan giữa các biến số. Giá trị tương quan nằm trong khoảng từ -1 đến 1, với 1 nghĩa là tương quan thuận hoàn hảo, -1 nghĩa là tương quan nghịch hoàn hảo, và 0 nghĩa là không có tương quan.
- ggplot(correlation_df, aes(x = Variable1, y = Variable2, fill = Correlation)) + ... Biểu đồ heatmap cho phép ta thấy rõ mối tương quan giữa các biến số một cách trực quan.
- Mối liên quan tích cực: B (tài sản ngắn hạn), H (các khoản phải thu ngắn hạn), I (phải thu ngắn hạn của khách hàng) và nhiều yếu tố khác có mối tương quan tích cực với nhau (các ô có màu đậm)
- Đồng nghĩa với việc các yếu tố này tỉ lệ thuận với nhau. Khi một yếu tố tăng thì các yếu tố khác cũng tăng theo.
- Mối liên quan không tích cực: BE (lợi nhuận phân phối sau thuế) và các yếu tố khác (Các ô có màu xám)
- Tức là: lợi nhuận được phân phối sau thuế không được cao.

2.2.3 Mối liên hệ giữa các biến

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.

Scatter_plot <- ggplot(data =) +
geom_point(mapping = aes(x = nhân tố liên quan 1, y = nhân tố liên quan 2)) +
Labs(title =)

Chúng tôi sử dụng biểu đồ phân tán để phân tích mối tương quan giữa lợi nhuận của hai mã chứng khoán.

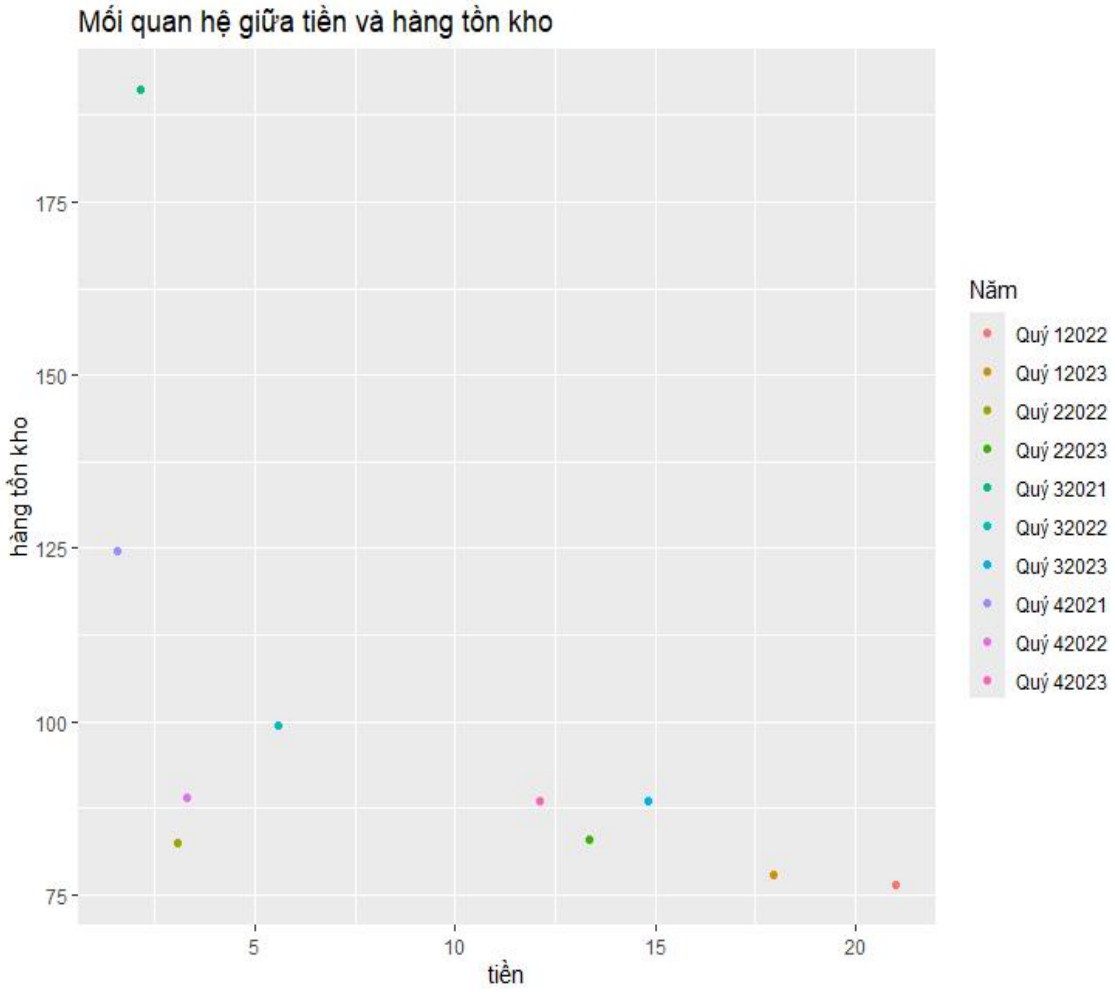
2.2.3.1 Mối tương quan giữa tiền và hàng tồn kho

Mối tương quan giữa tiền mặt (cash) và hàng tồn kho (inventory) phản ánh cách quản lý hàng tồn kho ảnh hưởng đến dòng tiền của doanh nghiệp. Quản lý hàng tồn kho hiệu quả giúp giảm chi phí lưu trữ và tăng dòng tiền, trong khi việc duy trì mức tồn kho cao có thể làm giảm lượng tiền mặt sẵn có.

```
#Mối liên hệ giữa tiền và hàng tồn kho
#scatter plot
ggplot(data = do_an1) + geom_point(mapping = aes(x = `D`, y = `O`, color = `Năm`)) + labs(x = "tiền", y = "hàng tồn kho" ,title = "Mối quan hệ giữa tiền và hàng tồn kho")
#ma trận tương quan
tien <- do_an1$D
hang_ton_kho <- do_an1$O
mlh <-data.frame(tien, hang_ton_kho)
raqMatrix2 <- cor(mlh)
#tính ma trận tương quan
round(raqMatrix2,5)
#kiểm định tính tương quan
cor.test(tien, hang_ton_kho)
```

A matrix: 2 × 2 of type dbl

	tien	hang_ton_kho
tien	1.0000	-0.5756
hang_ton_kho	-0.5756	1.0000



Hình 2.2 Mối tương quan giữa tiền và hàng tồn kho

Nhận xét

- Biểu đồ phân tán (Scatter plot): Biểu đồ này cho thấy mối quan hệ giữa tiền và hàng tồn kho. Nếu có một mối quan hệ tuyến tính rõ ràng giữa hai biến, điều này có thể cho thấy rằng việc tăng hoặc giảm số tiền có thể ảnh hưởng trực tiếp đến số lượng hàng tồn kho.
- Ma trận tương quan: Ma trận tương quan cung cấp một con số cụ thể để đo lường mức độ mà tiền và hàng tồn kho liên quan đến nhau. Điều này có thể giúp doanh nghiệp hiểu rõ hơn về mối quan hệ này và có thể sử dụng thông tin này để đưa ra quyết định về cách quản lý tài chính và hàng tồn kho.
- Kiểm định tương quan: Kiểm định tương quan giúp xác định liệu mối quan hệ tuyến tính giữa tiền và hàng tồn kho có ý nghĩa thống kê hay không. Nếu

kết quả kiểm định cho thấy mối quan hệ này có ý nghĩa, doanh nghiệp có thể tin tưởng rằng việc thay đổi số tiền sẽ có ảnh hưởng đến số lượng hàng tồn kho.

2.2.3.2 Mối tương quan giữa tài sản ngắn hạn và tài sản dài hạn

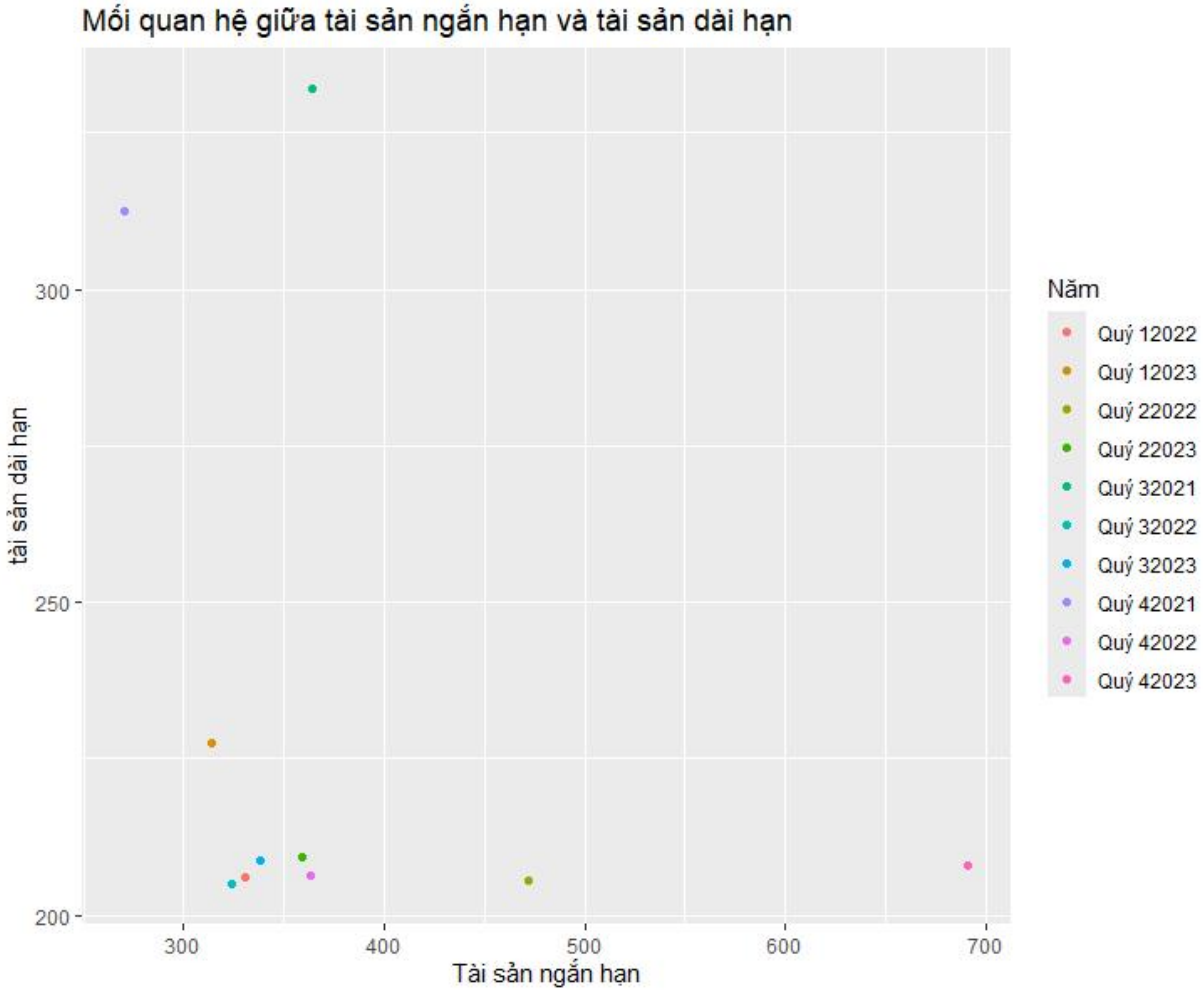
Mối tương quan giữa tài sản ngắn hạn (current assets) và tài sản dài hạn (non-current assets) phản ánh cách quản lý và đầu tư của doanh nghiệp. Tài sản ngắn hạn bao gồm các tài sản có thể chuyển đổi thành tiền mặt trong vòng một năm, như tiền mặt, các khoản phải thu và hàng tồn kho. Tài sản dài hạn là các tài sản có thời gian sử dụng trên một năm, như tài sản cố định và đầu tư dài hạn. Theo đó phân tích mối tương quan này chúng ta xây dựng biểu đồ dựa trên chương trình sau

```
#Mối liên hệ giữa tài sản ngắn hạn và tài sản dài hạn
library(ggplot2)
#biểu đồ scatter plot
ggplot(data = do_an1) + geom_point(mapping = aes(x = `B`, y = `U`, color = `Năm`)) + labs(x = "Tài sản ngắn hạn", y= "tài sản dài hạn" ,title = "Mối quan hệ giữa tài sản ngắn hạn và tài sản dài hạn")
#ma trận tương quan
install.packages("corrplot")
library(corrplot)
tai_san_ngan_han <- do_an1$B
tai_san_dai_han <- do_an1$U
doan <- data.frame(tai_san_dai_han, tai_san_ngan_han)
raqMatrix <- cor(doan)
#tính ma trận tương quan
round(raqMatrix,5)
#kiểm định tính tương quan
cor.test(tai_san_ngan_han, tai_san_dai_han)
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

A matrix: 2 × 2 of type dbl

	tai_san_dai_han	tai_san_ngan_han
tai_san_dai_han	1.00000	-0.29411
tai_san_ngan_han	-0.29411	1.00000



Hình 2.3 Mối tương quan giữa tài sản dài hạn và tài sản ngắn hạn

Nhận xét:

Biểu đồ này cho thấy mối quan hệ giữa tài sản ngắn hạn và tài sản dài hạn. Nếu có một mối quan hệ tuyến tính rõ ràng giữa hai biến, điều này có thể cho thấy rằng việc tăng hoặc giảm tài sản ngắn hạn có thể ảnh hưởng trực tiếp đến tài sản dài hạn.

- Ma trận tương quan: Ma trận tương quan cung cấp một con số cụ thể để đo lường mức độ mà tài sản ngắn hạn và tài sản dài hạn liên quan đến nhau. Điều này có thể giúp doanh nghiệp hiểu rõ hơn về mối quan hệ này và có thể sử dụng thông tin này để đưa ra quyết định về cách quản lý tài chính.
- Kiểm định tương quan: Kiểm định tương quan giúp xác định liệu mối quan hệ tuyến tính giữa tài sản ngắn hạn và tài sản dài hạn có ý nghĩa thống kê hay không. Nếu kết quả kiểm định cho thấy mối quan hệ này có ý nghĩa, doanh nghiệp có thể tin tưởng rằng việc thay đổi tài sản ngắn hạn sẽ có ảnh hưởng đến tài sản dài hạn.
- Ta có thể thấy tỉ lệ tài sản ngắn hạn quá cao so với tài sản dài hạn, điều này có thể cho thấy doanh nghiệp đang gặp khó khăn trong việc đầu tư vào tài sản cố định để mở rộng hoạt động kinh doanh

Kết luận

Tính ma trận tương quan giữa các biến khi phân tích dữ liệu chứng khoán doanh nghiệp là một bước quan trọng, cung cấp thông tin về các mối quan hệ giữa các biến số khác nhau trong dữ liệu tài chính. Chương này vừa giúp người đọc hiểu được mối quan hệ giữa các loại tài sản với nhau, nâng cao mức độ quản lý rủi ro và phát hiện bất thường cách nhanh nhất.

2.3 THỐNG KÊ MÔ TẢ

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

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

```
data_summary <- summary(data)
print(data_summary)
```

```
     Năm                 B               C               D
 Length:10          Min.   :270.5   Min.   : 1.564   Min.   : 1.564
 Class :character   1st Qu.:325.2   1st Qu.: 4.002   1st Qu.: 3.126
 Mode  :character   Median :348.7   Median :12.732   Median : 8.842
                    Mean   :382.6   Mean   :15.442   Mean   : 9.491
                    3rd Qu.:363.8   3rd Qu.:22.566   3rd Qu.:14.466
                    Max.   :691.3   Max.   :42.133   Max.   :21.021
       E                F               H               I
 Min.   :  0.00   Min.   : 0.00   Min.   : 85.97   Min.   : 61.30
 1st Qu.: 45.88   1st Qu.:42.83   1st Qu.:130.40   1st Qu.: 87.87
 Median : 48.77   Median :50.96   Median :176.37   Median : 93.14
 Mean   : 49.63   Mean   :46.72   Mean   :214.63   Mean   :153.64
 3rd Qu.: 50.23   3rd Qu.:57.69   3rd Qu.:221.86   3rd Qu.:143.13
 Max.   :108.31   Max.   :64.34   Max.   :545.70   Max.   :469.07
       J               M               O               Q
 Min.   : 8.498   Min.   : 1.263   Min.   : 76.37   Min.   :  1.203
 1st Qu.:16.135   1st Qu.:15.932   1st Qu.: 82.71   1st Qu.:  1.795
 Median :46.694   Median :34.956   Median : 88.60   Median :  9.361
 Mean   :42.091   Mean   :30.353   Mean   :100.14   Mean   :222.248
 3rd Qu.:64.182   3rd Qu.:37.952   3rd Qu.: 96.87   3rd Qu.:330.000
 Max.   :82.389   Max.   :64.271   Max.   :191.27   Max.   :896.000
       R               S               U              AB
```

Từ đó chọn dữ liệu cho năm 2022 và 2023

```
      data_subset <- subset(do_an1, Năm %in% c("Quý 12022", "Quý 22022",     "Quý 32022", "Quý 42022", "Quý 12023", "Quý 22023", "Quý 32023",     "Quý 42023"))
      data_subset
```

- Subset(do_an1, Năm %in% c("Quý 12022", "Quý 22022", "Quý 32022", "Quý 42022", "Quý 12023", "Quý 22023", "Quý 32023", "Quý 42023")): Hàm này lọc ra các dòng dữ liệu trong do_an1 mà cột Năm có giá trị nằm trong danh sách các quý của năm 2022 và 2023. Kết quả lọc được lưu trữ trong data_subset.
- data_subset: Đây là kết quả của việc lọc dữ liệu, chứa các dòng tương ứng với các quý của năm 2022 và 2023.

```
 Biến đổi cột Năm thành factor để đảm bảo thứ tự hiển thị đúng trên biểu đồ
      do_an1$Năm <- factor(do_an1$Năm, levels = c("Quý 12022", "Quý 22022", "Quý 32022", "Quý 42022","Quý 12023", "Quý 22023", "Quý 32023", "Quý 42023"))
      do_an1$Năm
```

Chương trình có kết quả

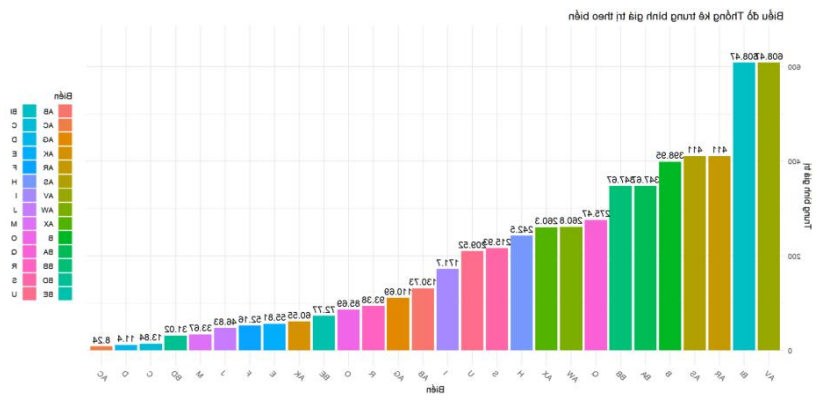
Quý 42023 · Quý 32023 · Quý 22023 · Quý 12023 · Quý 42022 · Quý 32022 · Quý 22022 · Quý 12022 ·

Chuyển đổi cột Năm trong dữ liệu do_an1 thành kiểu dữ liệu factor, với các mức (levels) được xác định rõ ràng theo thứ tự của các quý từ năm 2022 đến năm 2023. Điều này đảm bảo rằng khi vẽ biểu đồ, các mức của cột Năm sẽ được hiển thị đúng thứ tự mong muốn bằng hàm factor (do_an1$Năm, levels = c("Quý 12022", "Quý 22022", "Quý 32022", "Quý 42022", "Quý 12023", "Quý 22023", "Quý 32023", "Quý 42023"))

Từ quá trình mô tả trên hướng đến tính toán giá trị trung bình của các biến trong dữ liệu và sắp xếp chúng theo thứ tự giảm dần của giá trị trung bình. Bằng cách này, người dùng có thể nhanh chóng nhận biết được các biến có giá trị trung bình cao nhất và thấp nhất trong dữ liệu.

2.3.1 Biểu đồ bar

```
ggplot(data_filtered, aes(x = reorder(Biến, -Giá_trị), y = Giá_trị, fill = Biến)) +
  geom_bar(stat = "summary", fun = "mean") +
  geom_text(stat = "summary", aes(label = round(..y.., 2)), vjust = -0.5) +  # Thêm giá trị lên cột
  labs(title = "Biểu đồ Thống kê trung bình giá trị theo biến",
       x = "Biến",
       y = "Trung bình giá trị") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Hình 2.4 biểu đồ thống kê giá trị trung bình theo các biến

Nhận xét: sử dụng biểu đồ bar để biểu diễn giá trị trung bình của các biến một cách trực quan. Giúp so sánh giá trị trung bình giữa các biến và hiểu rõ hơn về phân phối của chúng trong dữ liệu. Thêm nhãn giá trị lên các thanh bar cũng giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn.

2.3.2 Biểu đồ cột

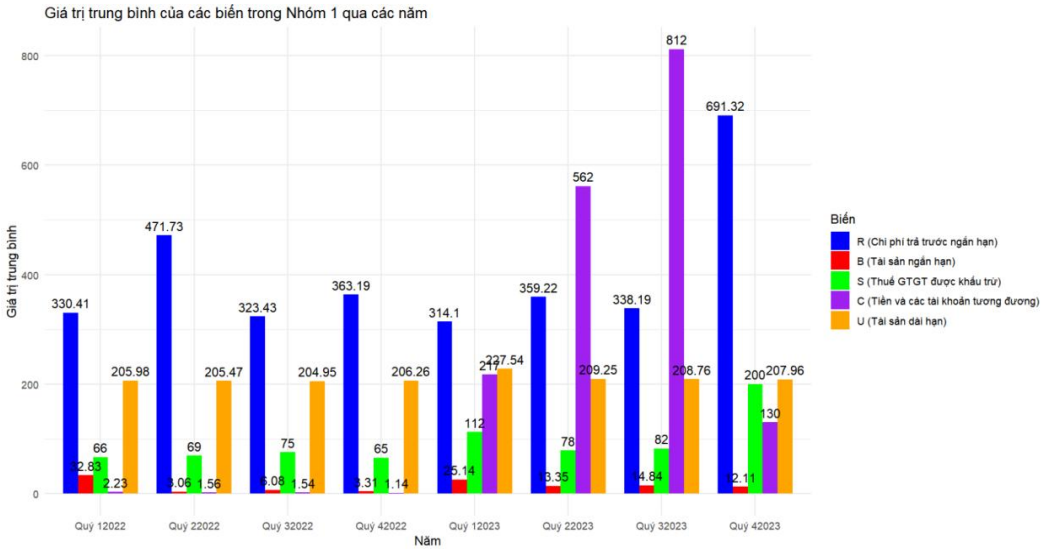
Tiến hành gộp nhóm 1, gồm tài sản ngắn hạn và nguồn vốn

```
group1 <- subset(data_filtered, Biến %in% c("R", "B", "S", "C",
"U"))
group1
# Tính giá trị trung bình cho từng biến và năm
mean_group1 <- aggregate(Giá_trị ~ Năm + Biến, data = group1,
FUN = mean)
mean_group1
```

Tạo một nhóm mới gồm các biến thuộc nhóm "Tài sản ngắn hạn" và "Nguồn vốn" từ tập dữ liệu đã lọc, và hiển thị nội dung của nhóm này để kiểm tra và tính toán giá trị trung bình của các biến trong nhóm group1 theo từng năm. Điều này giúp hiểu rõ hơn về xu hướng và biến động của các biến trong nhóm "Tài sản ngắn hạn" và "Nguồn vốn" qua các năm, cung cấp thông tin quan trọng để đánh giá và ra quyết định về tình hình tài chính của tổ chức hoặc doanh nghiệp.

```
# Vẽ biểu đồ cột
ggplot(data = mean_group1, aes(x = Năm, y = Giá_trị, fill =
Biến)) +
  geom_col(position = "dodge", width = 0.8) +
  labs(title = "Giá trị trung bình của các biến trong Nhóm 1
qua các năm",
       x = "Năm",
       y = "Giá trị trung bình") +
     geom_text(aes(label = round(Giá_trị, 2)), position =
position_dodge(width = 0.8), vjust = -0.5) +
```

```
    scale_fill_manual(values  =  c("blue",  "red",  "green",
"purple", "orange"),
                      labels = c("R (Chi phí trả trước ngắn hạn)",
                                 "B (Tài sản ngắn hạn)",
                                 "S (Thuế GTGT được khấu trừ)",
                                     "C (Tiền và các tài khoản
tương đương)",
                                 "U (Tài sản dài hạn)")) +
  theme_minimal()
```



Hình 2.5 giá trị trung bình của 5 biến trong một nhóm qua từng quý

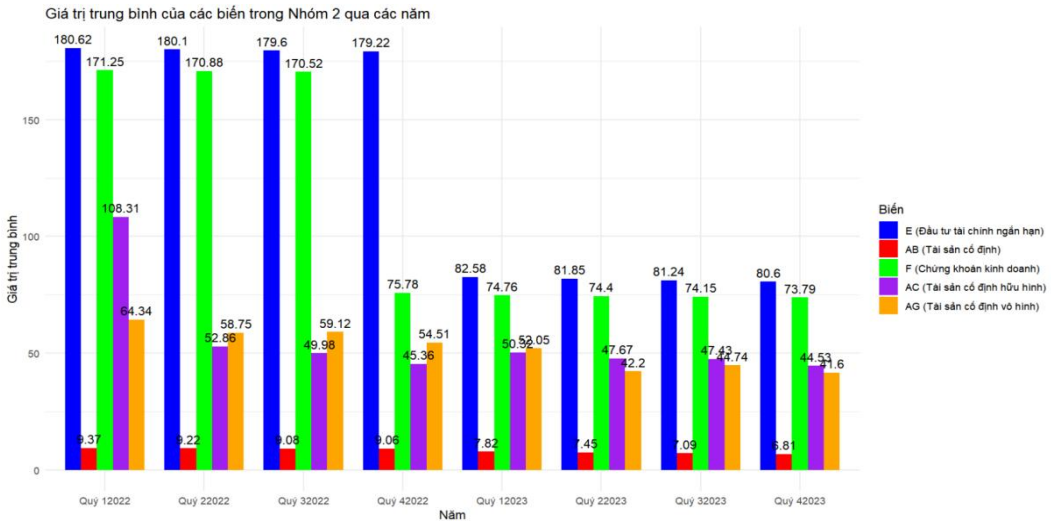
Trình bày thông tin một cách trực quan và dễ hiểu bằng cách sử dụng biểu đồ cột để biểu diễn giá trị trung bình của các biến trong nhóm 1 qua các năm. Biểu đồ cột giúp so sánh giá trị trung bình của các biến, cho phép người đọc nhận biết xu hướng và biến động của chúng qua thời gian. Thêm vào đó, việc đặt nhãn giá trị trên các cột cũng giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn. Điều này giúp người sử dụng nắm bắt được các thông tin chính, từ đó hỗ trợ trong việc ra quyết định và phân tích dữ liệu một cách hiệu quả.

Gộp nhóm 2

```
Gộp nhóm 2: Đầu tư và Chứng khoán
group2 <- subset(data_filtered, Biến %in% c("E", "AB", "F", "AC",
"AG"))
group2
# Tính giá trị trung bình cho từng biến và năm
mean_group2 <- aggregate(Giá_trị ~ Năm + Biến, data = group2,
FUN = mean)
mean_group
```

Từ đó tính toán giá trị trung bình cho từng biến và năm cung cấp thông tin về giá trị trung bình của các biến liên quan đến "Đầu tư và Chứng khoán" qua các năm, giúp hiểu rõ hơn về xu hướng và biến động của chúng. Điều này có thể hỗ trợ quyết định và phân tích về mặt tài chính của tổ chức hoặc doanh nghiệp.

```
# Vẽ biểu đồ cột
ggplot(data = mean_group2, aes(x = Năm, y = Giá_trị, fill = Biến)) +
  geom_col(position = "dodge", width = 0.8) +
  labs(title = "Giá trị trung bình của các biến trong Nhóm 2 qua các
năm",
       x = "Năm",
       y = "Giá trị trung bình") +
      geom_text(aes(label = round(Giá_trị, 2)), position =
position_dodge(width = 0.8), vjust = -0.5) +
    scale_fill_manual(values = c("blue", "red", "green", "purple",
"orange"),
                      labels = c("E (Đầu tư tài chính ngắn hạn)",
                                 "AB (Tài sản cố định)",
                                 "F (Chứng khoán kinh doanh)",
                                 "AC (Tài sản cố định hữu hình)",
                                 "AG (Tài sản cố định vô hình)")) +
  theme_minimal()
```



Hình 2.6 thống kê giá trị trung bình của 5 biến trong nhóm 2 qua từng quý

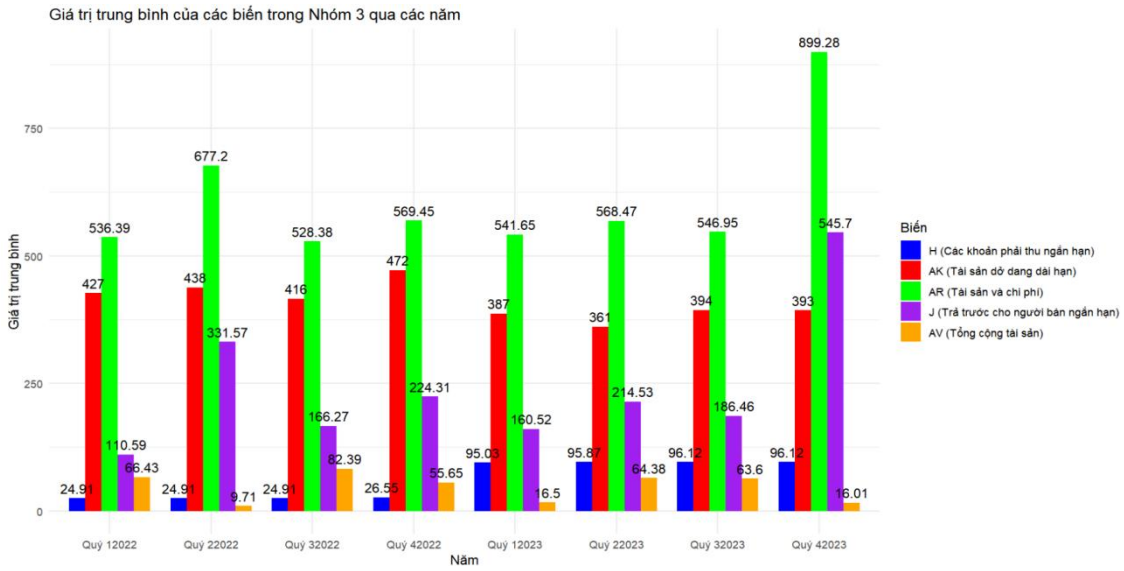
Chương trình vẽ một biểu đồ cột để trực quan hóa giá trị trung bình của các biến trong nhóm 2 qua các năm, so sánh giá trị trung bình của các biến và nhìn ra xu hướng hoặc biến động của chúng qua thời gian. Thêm vào đó, nhãn giá trị trên các cột giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn.

Gộp nhóm 3

```
# Gộp nhóm 3: Các khoản phải thu và Hàng tồn kho
group3 <- subset(data_filtered, Biến %in% c("H", "AK", "AR", "J", "AV"))
group3
# Tính giá trị trung bình cho từng biến và năm
mean_group3 <- aggregate(Giá_trị ~ Năm + Biến, data = group3, FUN = mean)
mean_group3
```

Tạo ra một tập dữ liệu mới chỉ chứa các dòng có giá trị trong cột "Biến" thuộc vào nhóm các biến "Các khoản phải thu và Hàng tồn kho", từ tập dữ liệu đã lọc trước đó. Điều này giúp tập trung vào các biến có liên quan đến các khoản phải thu và hàng tồn kho, tạo điều kiện thuận lợi cho việc phân tích và so sánh các chỉ số liên quan.Cung cấp thông tin về giá trị trung bình của các biến trong nhóm "Các khoản phải thu và Hàng tồn kho" qua các năm, giúp hiểu rõ hơn về xu hướng và biến động của chúng. Điều này có thể hỗ trợ quyết định và phân tích dữ liệu một cách hiệu quả.

```
# Vẽ biểu đồ cột
ggplot(data = mean_group3, aes(x = Năm, y = Giá_trị, fill = Biến)) +
  geom_col(position = "dodge", width = 0.8) +
  labs(title = "Giá trị trung bình của các biến trong Nhóm 3 qua các năm",
       x = "Năm",
       y = "Giá trị trung bình") +
      geom_text(aes(label = round(Giá_trị, 2)), position = position_dodge(width = 0.8), vjust = -0.5) +
    scale_fill_manual(values = c("blue", "red", "green", "purple", "orange"),
                      labels = c("H (Các khoản phải thu ngắn hạn)",
                                 "AK (Tài sản dở dang dài hạn)",
                                 "AR (Tài sản và chi phí)",
                                 "J (Trả trước cho người bán ngắn hạn)",
```



Hình 2.7 Giá trị trung bình của 5 biến trong nhóm 3 qua từng quý

Tương tự nhận xét trên về biểu đồ cột

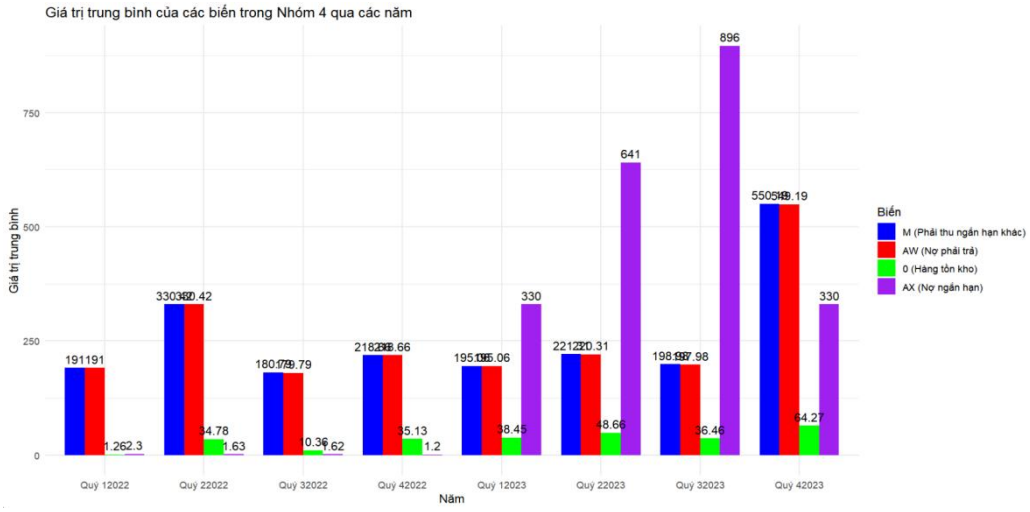
Gộp nhóm 4

```
# Gộp nhóm 4: Nợ và Tài sản ngắn hạn khác
group4 <- subset(data_filtered, Biến %in% c("M", "AW", "0", "AX", "Q"))
group4
# Tính giá trị trung bình cho từng biến và năm
```

```
mean_group4 <- aggregate(Giá_trị ~ Năm + Biến, data = group4, FUN =
mean)
mean_group4
```

Tạo ra một tập dữ liệu mới chỉ chứa các dòng có giá trị trong cột Biến thuộc vào nhóm các biến "Nợ và Tài sản ngắn hạn khác", từ tập dữ liệu đã lọc trước đó. Điều này giúp tập trung vào các biến có liên quan đến nợ và các loại tài sản ngắn hạn khác nhau, tạo điều kiện thuận lợi cho việc phân tích và so sánh các chỉ số liên quan. Cung cấp thông tin chi tiết về giá trị trung bình của các biến trong nhóm "Nợ và Tài sản ngắn hạn khác" qua các năm. Thông qua việc tính toán giá trị trung bình, người dùng có thể hiểu rõ hơn về xu hướng và biến động của các biến này qua thời gian, từ đó hỗ trợ trong việc đánh giá tình hình tài chính và ra quyết định.

```
# Vẽ biểu đồ cột
ggplot(data = mean_group4, aes(x = Năm, y = Giá_trị, fill = Biến)) +
  geom_col(position = "dodge", width = 0.8) +
  labs(title = "Giá trị trung bình của các biến trong Nhóm 4 qua các
năm",
       x = "Năm",
       y = "Giá trị trung bình") +
      geom_text(aes(label = round(Giá_trị, 2)), position =
position_dodge(width = 0.8), vjust = -0.5) +
   scale_fill_manual(values = c("blue", "red", "green", "purple",
"orange"),
                     labels = c("M (Phải thu ngắn hạn khác)",
                                "AW (Nợ phải trả)",
                                "0 (Hàng tồn kho)",
                                "AX (Nợ ngắn hạn)",
                                "Q (Tài sản ngắn hạn khác)")) +
  theme_minimal()
```



Hình 2.8 Giá trị trung bình của 4 biến trong nhóm 4 qua từng quý

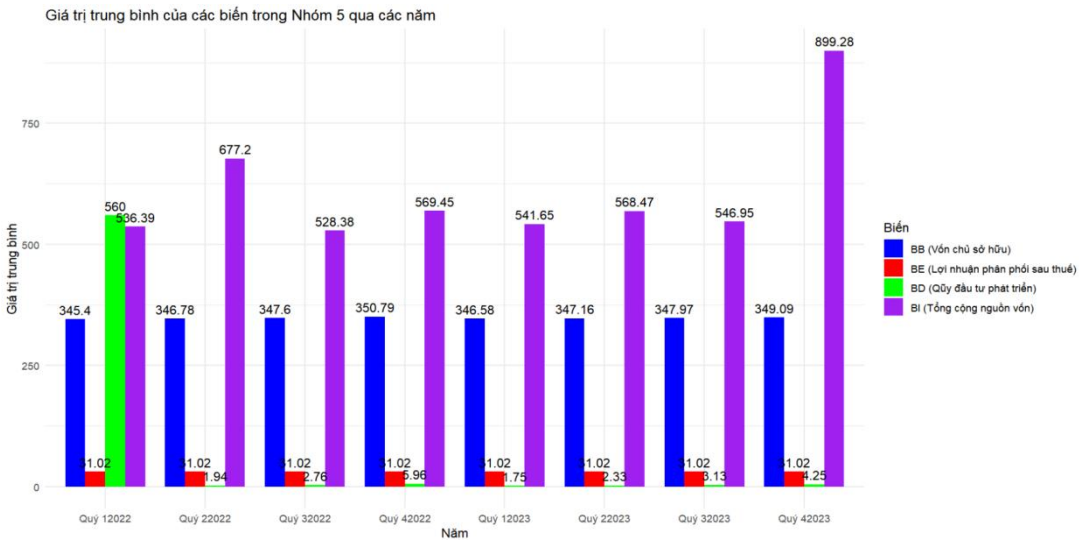
Gộp nhóm 5

```
# Gộp nhóm 5: Vốn chủ sở hữu và Qũy đầu tư
group5 <- subset(data_filtered, Biến %in% c("BB", "BE", "BD", "BI"))
group5
# Tính giá trị trung bình cho từng biến và năm
mean_group5 <- aggregate(Giá_trị ~ Năm + Biến, data = group5,
FUN = mean)
mean_group5
```

Tạo ra một tập dữ liệu mới chỉ chứa các dòng có giá trị trong cột "Biến" thuộc vào nhóm các biến "Vốn chủ sở hữu và Quỹ đầu tư", từ tập dữ liệu đã lọc trước đó. Điều này giúp tập trung vào các biến liên quan đến vốn chủ sở hữu và quỹ đầu tư, tạo điều kiện thuận lợi cho việc phân tích và so sánh các chỉ số liên quan. Cung cấp thông tin về giá trị trung bình của các biến trong nhóm "Vốn chủ sở hữu và Quỹ đầu tư" qua các năm, giúp hiểu rõ hơn về xu hướng và biến động của chúng. Điều này có thể hỗ trợ quyết định và phân tích dữ liệu một cách hiệu quả.

```
# Vẽ biểu đồ cột
ggplot(data = mean_group5, aes(x = Năm, y = Giá_trị, fill = Biến)) +
  geom_col(position = "dodge", width = 0.8) +
```

```
  labs(title = "Giá trị trung bình của các biến trong Nhóm 5 qua các năm",
       x = "Năm",
       y = "Giá trị trung bình") +
      geom_text(aes(label = round(Giá_trị, 2)), position = position_dodge(width = 0.8), vjust = -0.5) +
  scale_fill_manual(values = c("blue", "red", "green", "purple"),
                    labels = c("BB (Vốn chủ sở hữu)",
                               "BE (Lợi nhuận phân phối sau thuế)",
                               "BD (Qũy đầu tư phát triển)",
                               "BI (Tổng cộng nguồn vốn)")) +
  theme_minimal()
```

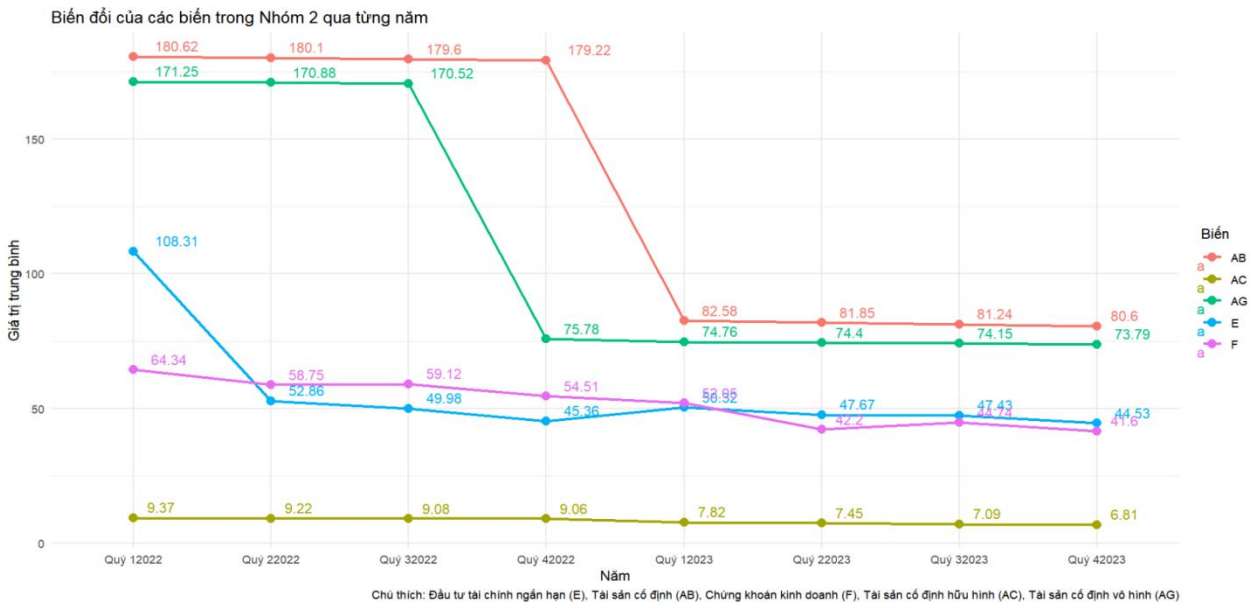


Hình 2.9 thống kê giá trị trung bình của 4 biến trong nhóm 5 qua từng quý

Trực quan hóa giá trị trung bình của các biến trong nhóm 5 qua các năm. Biểu đồ giúp so sánh giá trị trung bình của các biến và nhìn ra xu hướng hoặc biến động của chúng qua thời gian. Các cột được phân biệt bằng màu sắc, và nhãn giá trị trên các cột giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn. Điều này giúp người xem nhận biết các biến quan trọng và hiểu rõ hơn về vai trò của chúng trong quá trình kinh doanh hoặc tài chính.

2.3.2 Biểu đồ đường

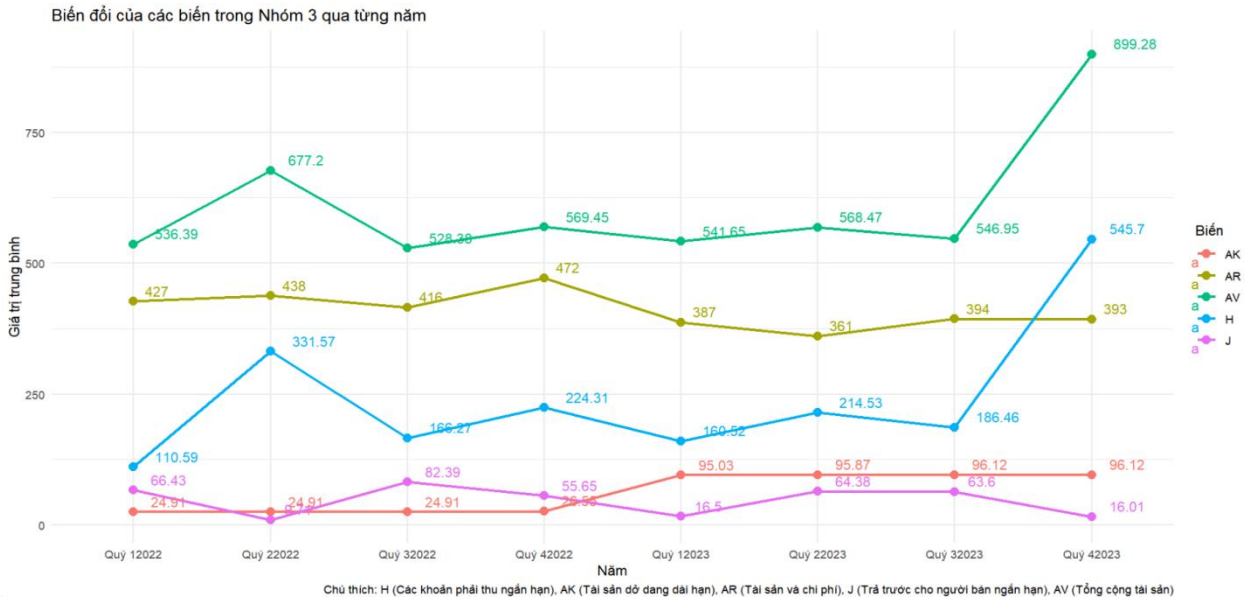
```
# Vẽ biểu đồ đường nhóm 2
ggplot(mean_group2, aes(x = factor(Năm), y = Giá_trị, color = Biến,
group = Biến)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
   geom_text(aes(label = round(Giá_trị, 2)), vjust = -0.5, hjust =
-0.5) +  # Thêm nhãn giá trị trung bình
  labs(title = "Biến đổi của các biến trong Nhóm 2 qua từng năm",
       x = "Năm",
       y = "Giá trị trung bình",
       color = "Biến",
        caption = "Chú thích: Đầu tư tài chính ngắn hạn (E), Tài
sản cố định (AB), Chứng khoán kinh doanh (F), Tài sản cố định hữu
hình (AC), Tài sản cố định vô hình (AG)") +
  theme_minimal()
```



Hình 2.10 biến đổi của các biến trong nhóm 2 qua từng quý

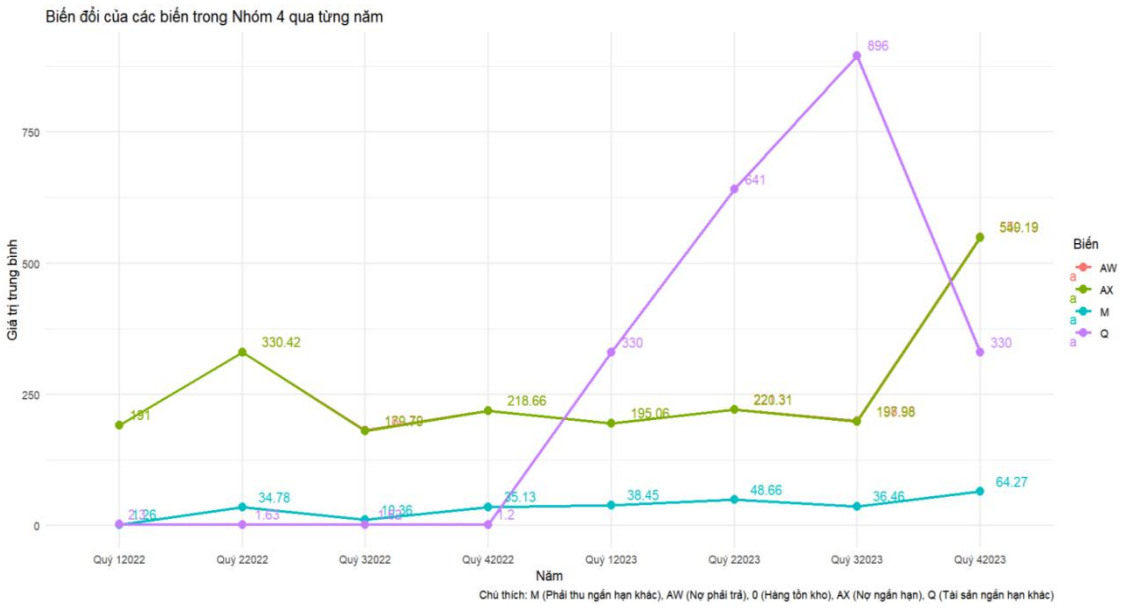
Trực quan hóa sự biến đổi của các biến trong nhóm 2 qua từng năm. Bằng cách này, bạn có thể dễ dàng nhận biết xu hướng tăng giảm hoặc biến động của các biến này qua thời gian. Các điểm dữ liệu trên đường và nhãn giá trị trung bình cũng được thêm vào để giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn. Mỗi màu sắc được gán cho một biến trong nhóm, giúp bạn dễ dàng phân biệt giữa các loại biến. Điều này giúp bạn hiểu rõ hơn về vai trò và ảnh hưởng của mỗi biến trong nhóm 2 đối với tổ chức hoặc doanh nghiệp.

```
# Vẽ biểu đồ đường nhóm 3
ggplot(mean_group3, aes(x = factor(Năm), y = Giá_trị, color = Biến,
group = Biến)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  geom_text(aes(label = round(Giá_trị, 2)), vjust = -0.5, hjust =
-0.5) +  # Thêm nhãn giá trị trung bình
  labs(title = "Biến đổi của các biến trong Nhóm 3 qua từng năm",
       x = "Năm",
       y = "Giá trị trung bình",
       color = "Biến",
       caption = "Chú thích: H (Các khoản phải thu ngắn hạn), AK
(Tài sản dở dang dài hạn), AR (Tài sản và chi phí), J (Trả trước
cho người bán ngắn hạn), AV (Tổng cộng tài sản)") +
                    theme_minimal()
```



Hình 2.11 biến đổi của các biến trong nhóm 3 qua từng quý

```
# Vẽ biểu đồ đường nhóm 4
ggplot(mean_group4, aes(x = factor(Năm), y = Giá_trị, color = Biến, group = Biến)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  geom_text(aes(label = round(Giá_trị, 2)), vjust = -0.5, hjust = -0.5) +  # Thêm nhãn giá trị trung bình
  labs(title = "Biến đổi của các biến trong Nhóm 4 qua từng năm",
       x = "Năm",
       y = "Giá trị trung bình",
       color = "Biến",
       caption = "Chú thích: M (Phải thu ngắn hạn khác), AW (Nợ phải trả), 0 (Hàng tồn kho), AX (Nợ ngắn hạn), Q (Tài sản ngắn hạn khác)") +
  theme_minimal()
```

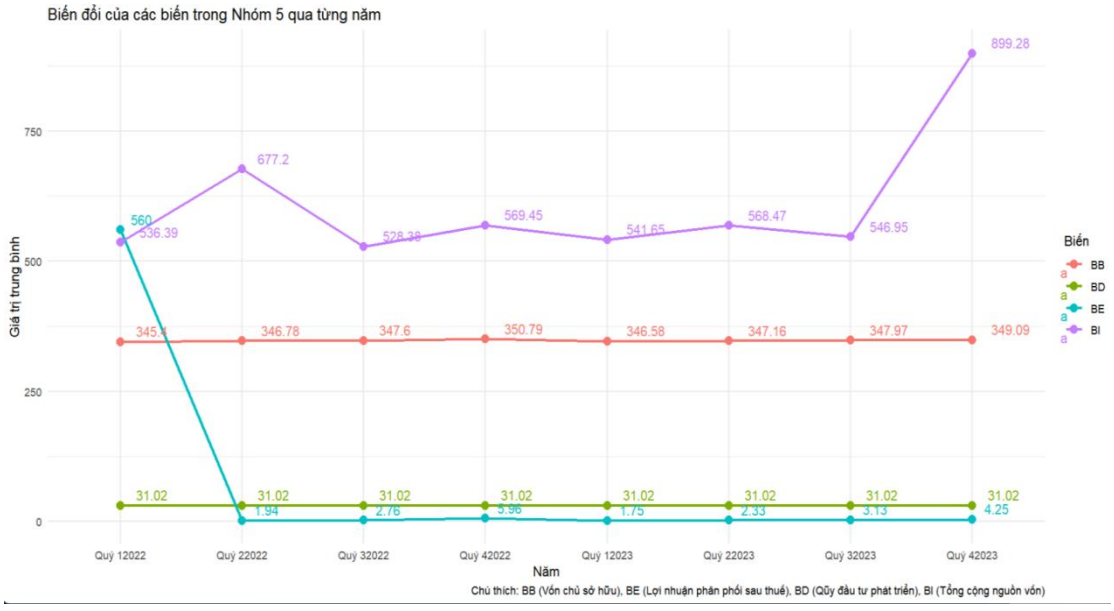


Hình 2.12 biến đổi của các biến trong nhóm 4 qua từng quý

Biểu đồ đường nhóm 4 trực quan hóa sự biến đổi của các biến trong nhóm 4 qua từng năm. Bằng cách này, bạn có thể dễ dàng nhận biết xu hướng tăng giảm

hoặc biến động của các biến này qua thời gian. Các điểm dữ liệu trên đường và nhãn giá trị trung bình cũng được thêm vào để giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn. Mỗi màu sắc được gán cho một biến trong nhóm, giúp bạn dễ dàng phân biệt giữa các loại biến. Điều này giúp bạn hiểu rõ hơn về vai trò và ảnh hưởng của mỗi biến trong nhóm 4 đối với tổ chức hoặc doanh nghiệp

```
# Vẽ biểu đồ đường nhóm 5
ggplot(mean_group5, aes(x = factor(Năm), y = Giá_trị, color = Biến, group = Biến)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  geom_text(aes(label = round(Giá_trị, 2)), vjust = -0.5, hjust = -0.5) +  # Thêm nhãn giá trị trung bình
  labs(title = "Biến đổi của các biến trong Nhóm 5 qua từng năm",
       x = "Năm",
       y = "Giá trị trung bình",
       color = "Biến",
       caption = "Chú thích: BB (Vốn chủ sở hữu), BE (Lợi nhuận phân phối sau thuế), BD (Qũy đầu tư phát triển), BI (Tổng cộng nguồn vốn)") +
  theme_minimal()
```



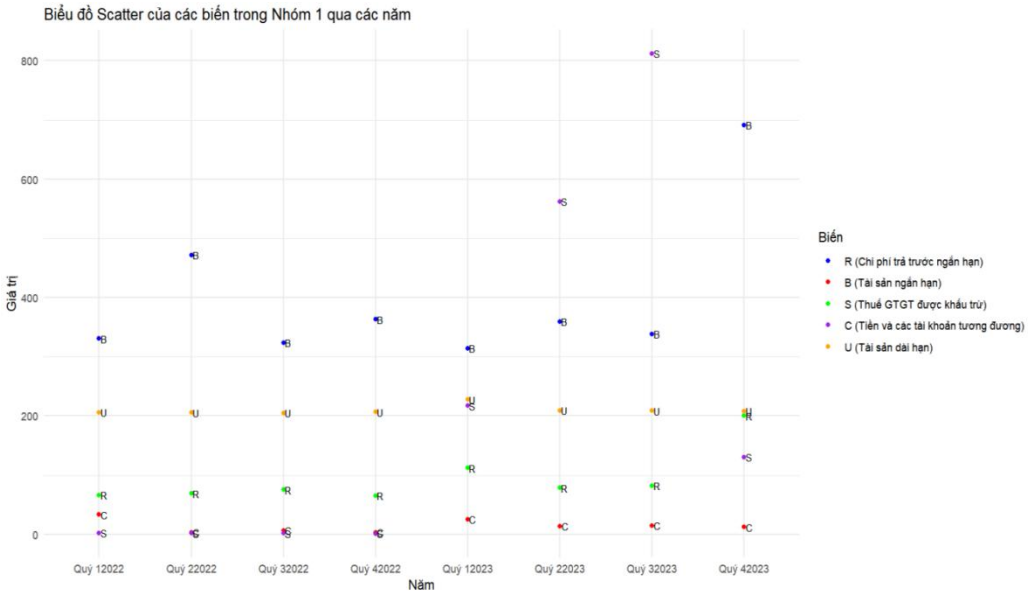
Hình 2.13. biến đổi của các biến trong nhóm 5 qua từng quý

Sự biến đổi của các biến trong nhóm 5 qua từng năm. Bằng cách này, bạn có thể dễ dàng nhận biết xu hướng tăng giảm hoặc biến động của các biến này qua thời gian. Các điểm dữ liệu trên đường và nhãn giá trị trung bình cũng được thêm vào để giúp làm cho thông tin trở nên dễ đọc và dễ hiểu hơn. Mỗi màu sắc được gán cho một biến trong nhóm, giúp bạn dễ dàng phân biệt giữa các loại biến. Điều này giúp bạn hiểu rõ hơn về vai trò và ảnh hưởng của mỗi biến trong nhóm 5 đối với tổ chức hoặc doanh nghiệp.

2.3.4 Biểu đồ scatter

```
# Vẽ biểu đồ scatter cho nhóm 1
ggplot(data = group1, aes(x = Năm, y = Giá_trị, color = Biến)) +
  geom_point() +
   labs(title = "Biểu đồ Scatter của các biến trong Nhóm 1 qua các
năm",
       x = "Năm",
       y = "Giá trị") +
   scale_color_manual(values = c("blue", "red", "green", "purple",
"orange"),
                      labels = legend_labels <- c("R (Chi phí trả
trước ngắn hạn)",
                                                  "B (Tài sản ngắn
hạn)",
                                                 "S (Thuế GTGT được
khấu trừ)",
                                                "C (Tiền và các tài
khoản tương đương)",
                                                   "U (Tài sản dài
hạn)")
  ) +

  geom_text(data = unique(group1[, c("Năm", "Giá_trị", "Biến")]),
            aes(label = Biến), hjust = -0.2, vjust = 0.5, size = 3,
color = "black") +
```

```
  theme_minimal()
```

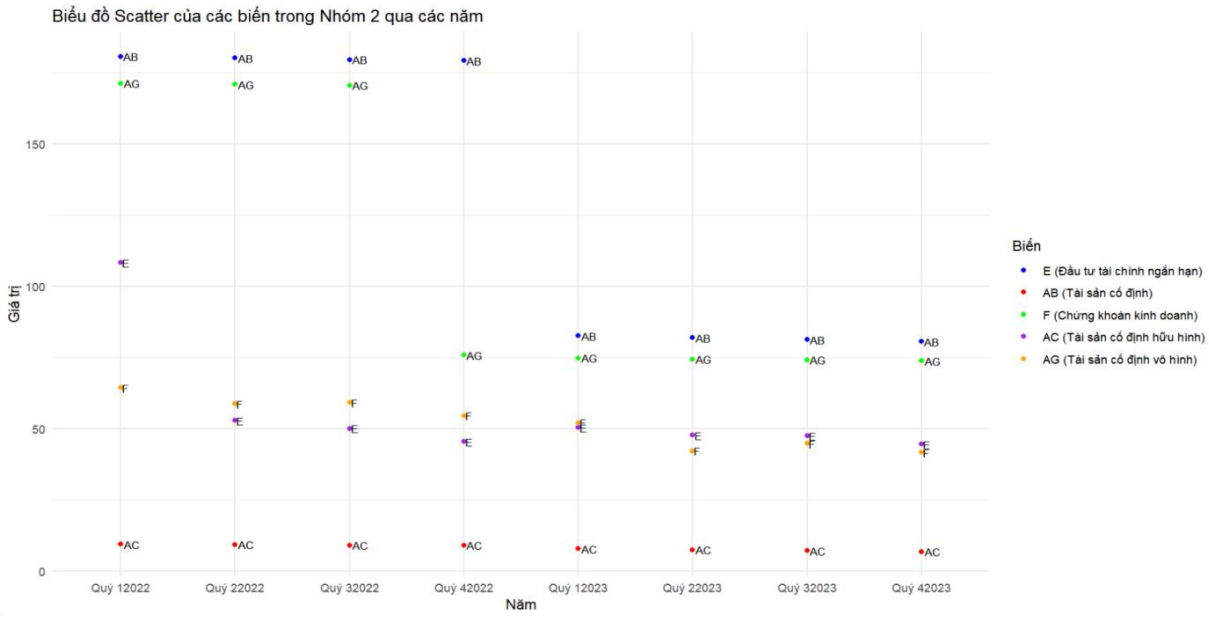


Hình 2.14 phân tán các biến trong nhóm 1 qua từng quý

Trực quan hóa giá trị của các biến trong nhóm 1 qua các năm. Điều này cho phép bạn nhận thấy các xu hướng, mối quan hệ hoặc sự phân tán của các biến theo thời gian. Mỗi điểm trên biểu đồ đại diện cho giá trị của một biến trong một năm cụ thể, với màu sắc khác nhau giúp phân biệt giữa các biến. Các nhãn biến được thêm vào gần các điểm dữ liệu để làm rõ hơn giá trị của từng biến. Biểu đồ này cung cấp cái nhìn chi tiết về sự phân bố và thay đổi của các biến trong nhóm 1, giúp xác định các xu hướng và bất thường qua các năm.

```
# Vẽ biểu đồ scatter nhóm 2
ggplot(data = group2, aes(x = Năm, y = Giá_trị, color = Biến)) +
  geom_point() +
  labs(title = "Biểu đồ Scatter của các biến trong Nhóm 2 qua các
năm",
       x = "Năm",
       y = "Giá trị") +
  scale_color_manual(values = c("blue", "red", "green", "purple",
"orange"),
```

```
                        labels = legend_labels <- c("E (Đầu tư tài
chính ngắn hạn)",
                                                    "AB (Tài sản cố
định)",
                                                    "F (Chứng khoán
kinh doanh)",
                                                    "AC (Tài sản cố
định hữu hình)",
                                                    "AG (Tài sản cố
định vô hình)")) +
  geom_text(data = unique(group2[, c("Năm", "Giá_trị", "Biến")]),
            aes(label = Biến), hjust = -0.2, vjust = 0.5, size = 3,
color = "black") +
  theme_minimal()
```



Hình 2.15 phân tán các biến trong nhóm 2 qua từng quý

Biểu đồ scatter của nhóm 2 cho chúng ta cái nhìn trực quan về sự phân bố và biến động của các giá trị thuộc nhóm đầu tư và chứng khoán trong doanh nghiệp qua các năm. Dưới đây là những điều mà biểu đồ này cho thấy về tình hình tài chính của doanh nghiệp

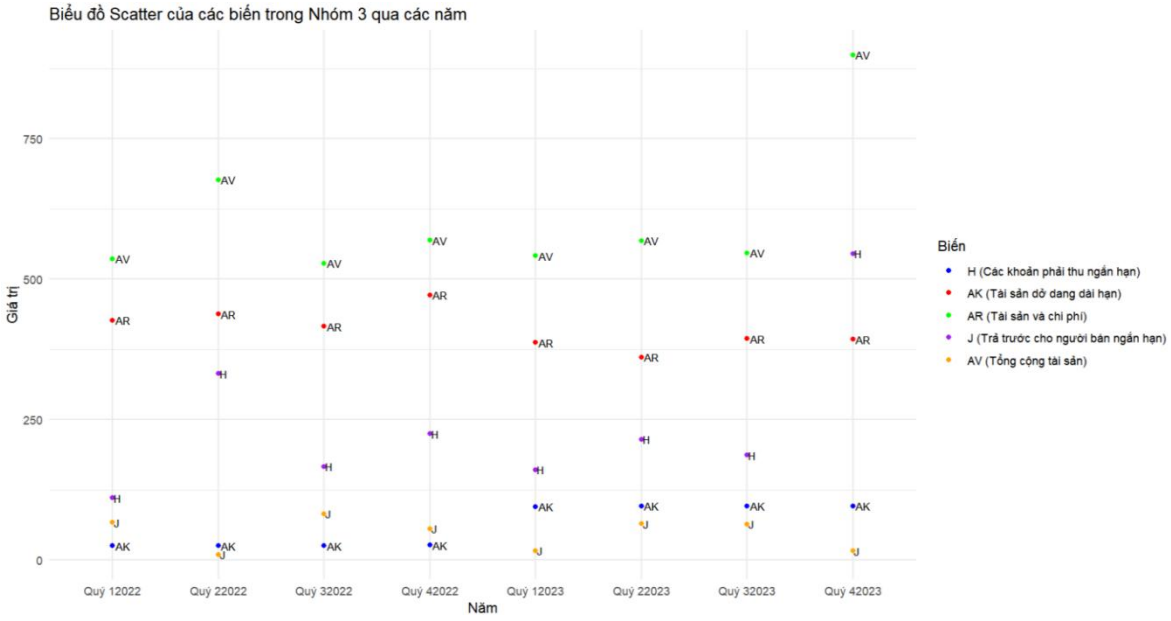
Phân bố giá trị qua các năm: Biểu đồ hiển thị cách mà giá trị của các biến như Đầu tư tài chính ngắn hạn (E), Tài sản cố định (AB), Chứng khoán kinh doanh (F), Tài sản cố định hữu hình (AC), và Tài sản cố định vô hình (AG) phân bố theo các năm. Điều này giúp xác định xu hướng tổng thể của từng loại tài sản hoặc khoản đầu tư trong doanh nghiệp.

So sánh giá trị giữa các biến: Sử dụng màu sắc khác nhau để biểu thị các biến khác nhau, biểu đồ giúp so sánh giá trị giữa các biến trong cùng một năm. Bạn có thể nhận thấy biến nào chiếm tỷ trọng lớn hơn hoặc nhỏ hơn trong từng năm cụ thể, từ đó hiểu rõ hơn về cấu trúc tài sản của doanh nghiệp.

Xu hướng biến động: Biểu đồ cung cấp thông tin về xu hướng tăng giảm của từng biến qua thời gian. Ví dụ, nếu giá trị đầu tư tài chính ngắn hạn (E) đang tăng lên qua các năm, điều này có thể cho thấy doanh nghiệp đang chú trọng vào các khoản đầu tư ngắn hạn. Ngược lại, nếu giá trị của tài sản cố định (AB, AC, AG) tăng, điều này có thể chỉ ra sự đầu tư vào tài sản dài hạn và hạ tầng.

```
# Vẽ biểu đồ scatter nhóm 3
ggplot(data = group3, aes(x = Năm, y = Giá_trị, color = Biến)) +
  geom_point() +
   labs(title = "Biểu đồ Scatter của các biến trong Nhóm 3 qua các năm",
        x = "Năm",
        y = "Giá trị") +
   scale_color_manual(values = c("blue", "red", "green", "purple", "orange"),
                      labels = legend_labels <- c("H (Các khoản phải thu ngắn hạn)",
                                                  "AK (Tài sản dở dang dài hạn)",
                                                   "AR (Tài sản và chi phí)",
                                                     "J (Trả trước cho người bán ngắn hạn)",
                                                    "AV (Tổng cộng tài sản)")) +
  geom_text(data = unique(group3[, c("Năm", "Giá_trị", "Biến")]),
```

```
            aes(label = Biến), hjust = -0.2, vjust = 0.5, size = 3,
color = "black") +
  theme_minimal()
```



Hình 2.16 phân tán các biến trong nhóm 3 qua từng quý

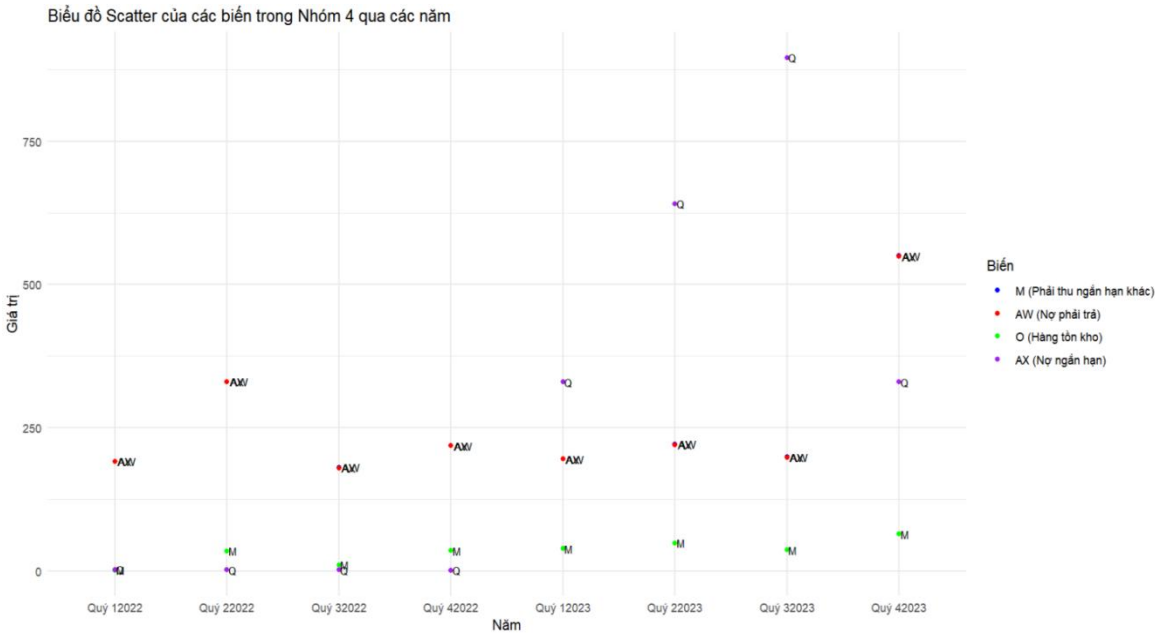
Phân bố giá trị theo thời gian: Biểu đồ hiển thị cách mà các giá trị của các biến như Các khoản phải thu ngắn hạn (H), Tài sản dở dang dài hạn (AK), Tài sản và chi phí (AR), Trả trước cho người bán ngắn hạn (J), và Tổng cộng tài sản (AV) phân bố qua các năm. Điều này giúp xác định các xu hướng dài hạn trong các loại tài sản và khoản phải thu của doanh nghiệp.

So sánh giá trị giữa các biến: Với việc sử dụng các màu sắc khác nhau cho mỗi biến, biểu đồ giúp bạn dễ dàng so sánh giá trị của từng biến trong từng năm. Bạn có thể nhận thấy biến nào có giá trị cao hơn hoặc thấp hơn trong một năm cụ thể, cung cấp cái nhìn về mức độ quan trọng và đóng góp của từng loại tài sản và khoản phải thu.

Xu hướng biến động: Biểu đồ cung cấp thông tin về xu hướng tăng hoặc giảm của từng biến theo thời gian. Ví dụ, nếu giá trị các khoản phải thu ngắn hạn (H) đang tăng dần qua các năm, điều này có thể cho thấy sự gia tăng trong các khoản tín dụng mà doanh nghiệp cấp cho khách hàng. Ngược lại, nếu giá trị tài

sản dở dang dài hạn (AK) giảm, điều này có thể phản ánh sự hoàn thành hoặc giảm đầu tư vào các dự án dài hạn.

```
# Vẽ biểu đồ scatter nhóm 4
ggplot(data = group4, aes(x = Năm, y = Giá_trị, color = Biến)) +
  geom_point() +
   labs(title = "Biểu đồ Scatter của các biến trong Nhóm 4 qua các năm",
        x = "Năm",
        y = "Giá trị") +
    scale_color_manual(values = c("blue", "red", "green", "purple", "orange"),
                       labels = legend_labels <- c("M (Phải thu ngắn hạn khác)",
                                                   "AW (Nợ phải trả)",
                                                   "O (Hàng tồn kho)",
                                                   "AX (Nợ ngắn hạn)",
                                                   "Q (Tài sản ngắn hạn khác)")) +
  geom_text(data = unique(group4[, c("Năm", "Giá_trị", "Biến")]),
            aes(label = Biến), hjust = -0.2, vjust = 0.5, size = 3, color = "black") +
  theme_minimal()
```



Hình 2.17 phân tán các biến trong nhóm 4 qua từng quý

Sự phân bố và biến động của các biến liên quan đến nợ và tài sản ngắn hạn khác trong doanh nghiệp qua các năm. Dưới đây là những điều mà biểu đồ này cho thấy:

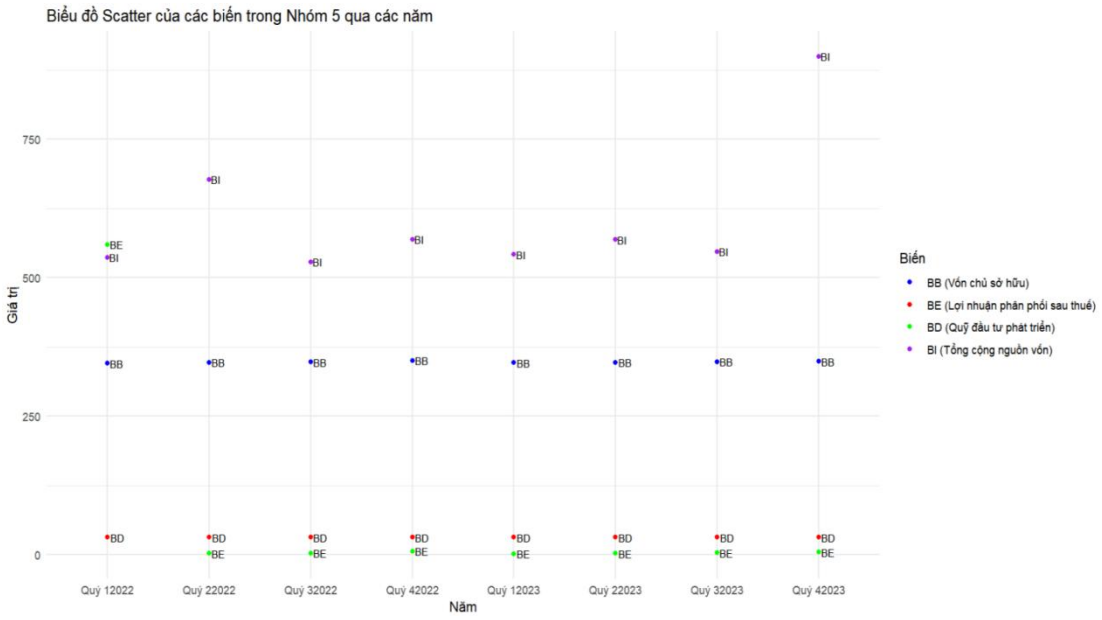
Phân bố giá trị theo thời gian: Biểu đồ hiển thị cách mà các giá trị của các biến như Phải thu ngắn hạn khác (M), Nợ phải trả (AW), Hàng tồn kho (O), Nợ ngắn hạn (AX), và Tài sản ngắn hạn khác (Q) phân bố qua các năm. Điều này giúp xác định xu hướng dài hạn trong các loại nợ và tài sản ngắn hạn khác của doanh nghiệp.

So sánh giá trị giữa các biến: Sử dụng các màu sắc khác nhau cho mỗi biến giúp dễ dàng so sánh giá trị của từng biến trong từng năm. Bạn có thể nhận thấy biến nào có giá trị cao hơn hoặc thấp hơn trong một năm cụ thể, cung cấp cái nhìn về mức độ quan trọng và đóng góp của từng loại nợ và tài sản ngắn hạn khác.

Xu hướng biến động: Biểu đồ cung cấp thông tin về xu hướng tăng hoặc giảm của từng biến theo thời gian. Ví dụ, nếu giá trị nợ phải trả (AW) đang tăng dần qua các năm, điều này có thể cho thấy sự gia tăng trong các khoản nợ của doanh nghiệp. Ngược lại, nếu giá trị hàng tồn kho (O) giảm, điều này có thể phản ánh sự giảm sản xuất hoặc hiệu quả trong việc quản lý hàng tồn kho.

```
# Vẽ biểu đồ scatter nhóm 5
ggplot(data = group5, aes(x = Năm, y = Giá_trị, color = Biến)) +
  geom_point() +
  labs(title = "Biểu đồ Scatter của các biến trong Nhóm 5 qua các
năm",
       x = "Năm",
       y = "Giá trị") +
  scale_color_manual(values = c("blue", "red", "green", "purple",
"orange"),
                     labels = legend_labels <- c("BB (Vốn chủ sở
hữu)",
                                                 "BE (Lợi nhuận
phân phối sau thuế)",
```

```
                                                    "BD (Quỹ đầu tư
phát triển)",
                                                     "BI (Tổng cộng
nguồn vốn)")) +
  geom_text(data = unique(group5[, c("Năm", "Giá_trị", "Biến")]),
            aes(label = Biến), hjust = -0.2, vjust = 0.5, size = 3,
color = "black") +
        theme_minimal()
```



Hình 2.18 phân tán các biến trong nhóm 5 qua từng quý

Phân bố giá trị theo thời gian: Biểu đồ hiển thị cách mà các giá trị của các biến như Vốn chủ sở hữu (BB), Lợi nhuận phân phối sau thuế (BE), Quỹ đầu tư phát triển (BD), và Tổng cộng nguồn vốn (BI) phân bố qua các năm. Điều này giúp xác định xu hướng dài hạn trong các thành phần vốn chủ sở hữu và quỹ đầu tư của doanh nghiệp.

So sánh giá trị giữa các biến: Sử dụng các màu sắc khác nhau cho mỗi biến giúp dễ dàng so sánh giá trị của từng biến trong từng năm. Bạn có thể nhận thấy biến nào có giá trị cao hơn hoặc thấp hơn trong một năm cụ thể, cung cấp cái nhìn về mức độ quan trọng và đóng góp của từng thành phần vốn và quỹ đầu tư.

Xu hướng biến động: Biểu đồ cung cấp thông tin về xu hướng tăng hoặc giảm của từng biến theo thời gian. Ví dụ, nếu giá trị lợi nhuận phân phối sau thuế (BE) đang tăng dần qua các năm, điều này có thể cho thấy doanh nghiệp đang tạo ra nhiều lợi nhuận hơn. Ngược lại, nếu quỹ đầu tư phát triển (BD) giảm, điều này có thể phản ánh sự giảm đầu tư vào phát triển của doanh nghiệp

CHƯƠNG 3: TRỰC QUAN HÓA DỮ LIỆU

3.1. TỔNG QUAN VỀ TRỰC QUAN HÓA DỮ LIỆU

Trực quan hóa dữ liệu là quá trình biến đổi dữ liệu thành các hình ảnh, biểu đồ, đồ thị hoặc bản đồ để hiểu và trình bày thông tin một cách trực quan và dễ hiểu. Mục tiêu chính của việc trực quan hóa dữ liệu là giúp người sử dụng dễ dàng nhận ra các mẫu, xu hướng, và mối quan hệ trong dữ liệu một cách nhanh chóng và hiệu quả.

Việc này giúp người đọc hiểu dữ liệu một cách nhanh chóng, dữ liệu trở nên trực quan hơn, giúp người xem dễ dàng hiểu thông tin và phát hiện mẫu một cách nhanh chóng. Phát hiện xu hướng và mối quan hệ người xem có thể dễ dàng nhận ra các xu hướng, mối quan hệ và biến động trong dữ liệu, giúp họ đưa ra các quyết định có logic hơn. Hơn hết là tạo ra câu chuyện từ dữ liệu từ đó giải thích và minh họa các thông tin phức tạp một cách đơn giản và sinh động. Đồng thời đưa ra dự đoán và phân tích tương lai.

Một số biểu đồ được sử dụng trong bài báo cáo:

- 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.
- Biểu đồ Heatmap: Thể hiện cường độ dữ liệu qua màu sắc

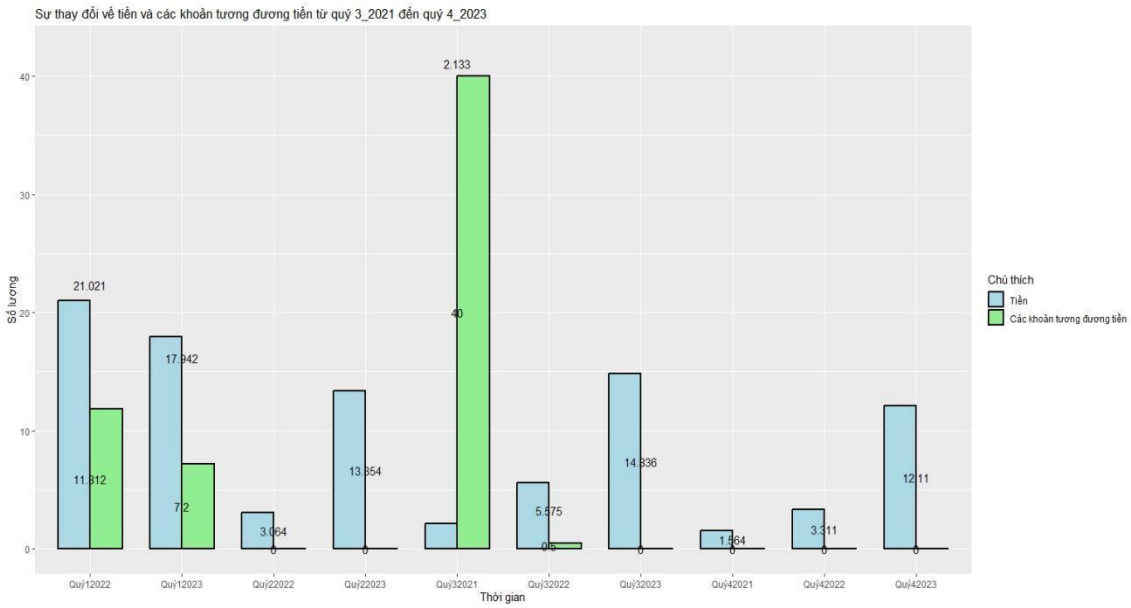
3.2. BIỂU ĐỒ TRỰC QUAN

3.2.1 Biểu đồ cột ghép

Biểu đồ này thể hiện sự thay đổi về tiền và các khoản tương đương tiền từ quý 3_2021 đến quý 4_2023. Biểu đồ được xây dựng theo chương trình sau:

```
# Vẽ biểu đồ thể hiện sự thay đổi về tiền và các khoản tương đương tiền từ
quý 3_2021 đến quý 4_2023
# Trích dữ liệu cột thời gian
time<-c(do_an1$A)
# Trích dữ liệu cột giá trị tiền
D<-c(do_an1$D)
# Tính giá trị các khoản tương đương tiền
result<-c(C-D)
data1<-data.frame(time=time, D=D, result=result)
data1_long <- tidyr::gather(data1, key = "variable", value = "value", -time)
# Vẽ biểu đồ cột ghép
ggplot(data1_long, aes(x = time, y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7) +
  geom_bar(stat = "identity", position = "dodge", width = 0.7, color =
"black", linewidth = 1) + # Thêm viền cho biểu đồ
  geom_text(aes(label = value), position = position_stack(vjust = 0.5),
color = "black") + # Thêm số liệu vào biểu đồ
  labs(title = "Sự thay đổi về tiền và các khoản tương đương tiền từ quý
3_2021 đến quý 4_2023",
       x = "Thời gian", y = "Số lượng") +
  scale_fill_manual(values = c("lightblue", "lightgreen"),
                    labels = c("Tiền", "Các khoản tương đương tiền"),
                    name = "Chú thích")
```

Từ đó đưa ra được hình ảnh trực quan



Hình 3.1: sự thay đổi về tiền và các khoản tương đương tiền từ quý 3_2021 đến quý 4_2023

Nhận xét

Qua đó cho thấy tổng số tiền và các khoản tương đương tiền giảm từ quý 3 năm 2021 đến quý 4 năm 2023 tiền giảm mạnh hơn các khoản tương đương tiền. Các khoản tiền: Giảm từ 21.021 tỷ USD xuống còn 13.854 tỷ USD, tương đương giảm 33,6%. Giảm mạnh nhất trong quý 2 năm 2022 (giảm 23,9%) và quý 4 năm 2023 (giảm 21,4%). Các khoản tương đương tiền: Giảm từ 17.942 tỷ USD xuống còn 14.836 tỷ USD, tương đương giảm 17,4%. Giảm mạnh nhất trong quý 4 năm 2023 (giảm 17,2%).

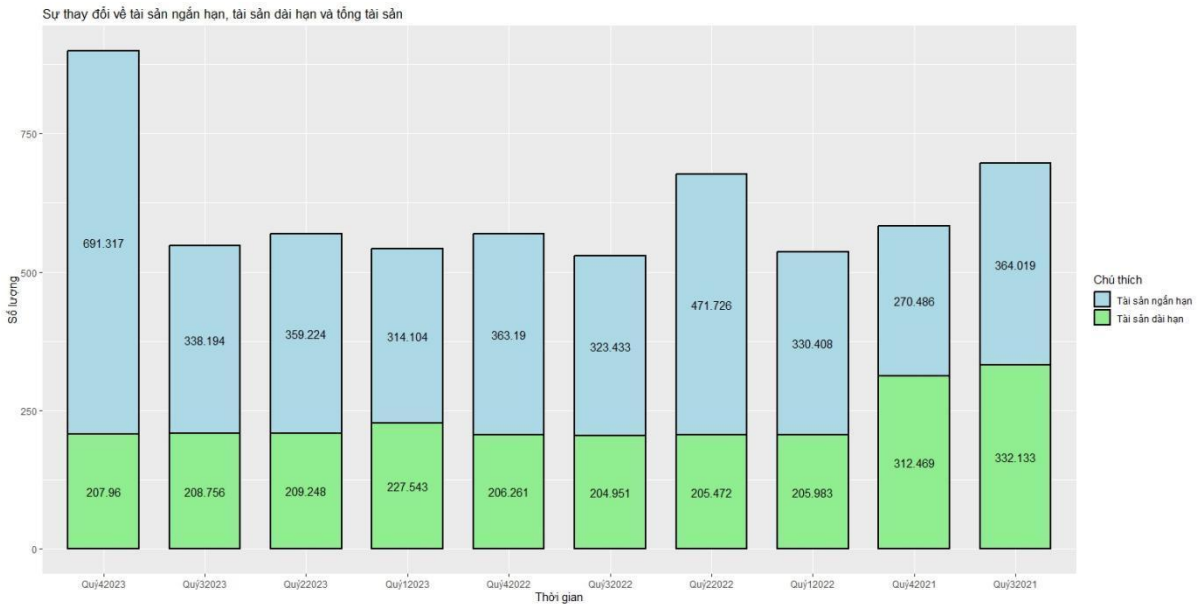
Biểu đồ cũng cho thấy sự sụt giảm đáng kể về tiền và các khoản tương đương tiền từ quý 3 năm 2021 đến quý 4 năm 2023. Sự sụt giảm này có thể do một số yếu tố, chẳng hạn như:

- Thị trường tài chính biến động: Thị trường tài chính biến động có thể dẫn đến thua lỗ đầu tư, từ đó dẫn đến giảm tiền mặt và các khoản tương đương tiền.
- Thay đổi tỷ giá hối đoái: Thay đổi tỷ giá hối đoái có thể ảnh hưởng đến giá trị của các khoản tương đương tiền bằng ngoại tệ.

3.2.2 Biểu đồ cột chồng.

Biểu đồ này thường được sử dụng để so sánh tỷ lệ phần trăm của các thành phần riêng lẻ trong một toàn bộ, thường là trong một chu kỳ thời gian cụ thể, như ngày, tuần, tháng hoặc năm. Đưa ra sự thay đổi về tài sản ngắn hạn, tài sản dài hạn và tổng tài sản. Chương trình phân tích như sau:

```
# Vẽ biểu đồ thể hiện sự thay đổi về tài sản ngắn hạn, tài sản dài hạn và
tổng tài sản
# Trích dữ liệu cột giá trị tài sản ngắn hạn
B <- c(do_an1$B)
# Trích dữ liệu cột giá trị tài sản dài hạn
U<-c(do_an1$U)
#Tạo dataframe
data8<-data.frame(time=time, B=B, U=U)
data8$time <- factor(data8$time, levels = time)
# Chuyển đổi dữ liệu sang dạng dài
data8_long <- tidyr::gather(data8, key = "variable", value = "value", -time)
# Vẽ biểu đồ cột chồng
ggplot(data8_long, aes(x = time, y = value, fill = variable)) +
    geom_bar(stat = "identity", position = "stack", width = 0.7) +
    geom_bar(stat = "identity", position = "stack", width = 0.7, color =
"black", linewidth = 1) + # Thêm viền cho biểu đồ
    geom_text(aes(label = value), position = position_stack(vjust = 0.5),
color = "black") + # Thêm số liệu vào biểu đồ
    labs(title = "Sự thay đổi về tài sản ngắn hạn, tài sản dài hạn và tổng
tài sản",
       x = "Thời gian", y = "Số lượng") +
    scale_fill_manual(values = c("lightblue", "lightgreen"),
                      labels = c("Tài sản ngắn hạn", "Tài sản dài hạn"),
                      name = "Chú thích")
```



Hình 3.2: sự thay đổi về tài sản ngắn hạn, tài sản dài hạn và tổng tài sản

Các loại tài sản xuất hiện là tài sản ngắn hạn, Tài sản dài hạn, Tổng tài sản.

Xu hướng chung: Cả ba loại tài sản đều có xu hướng giảm dần từ quý 4 năm 2023 đến quý 1 năm 2024. Tài sản dài hạn luôn cao hơn tài sản ngắn hạn trong suốt 4 quý được khảo sát. Tổng tài sản giảm từ 750 tỷ USD xuống còn 691,317 tỷ USD, tương đương giảm 7,75%.

- Tài sản ngắn hạn: Giảm từ 364,019 tỷ USD xuống còn 312,469 tỷ USD, tương đương giảm 14,2%. Giảm mạnh nhất trong quý 1 năm 2024 (giảm 16,5%).
- Tài sản dài hạn: Giảm từ 383,19 tỷ USD xuống còn 378,848 tỷ USD, tương đương giảm 1,1%. Giảm mạnh nhất trong quý 1 năm 2024 (giảm 1,4%).
- Tổng tài sản: Giảm từ 750 tỷ USD xuống còn 691,317 tỷ USD, tương đương giảm 7,75%. Giảm mạnh nhất trong quý 1 năm 2024 (giảm 7,9%).

So sánh các quý:

- Tỷ lệ tài sản ngắn hạn so với tổng tài sản: Giảm từ 48,54% xuống còn 45,19%, tương đương giảm 3,35%.

- Tỷ lệ tài sản dài hạn so với tổng tài sản: Giảm từ 51,46% xuống còn 54,81%, tương đương tăng 3,35%.

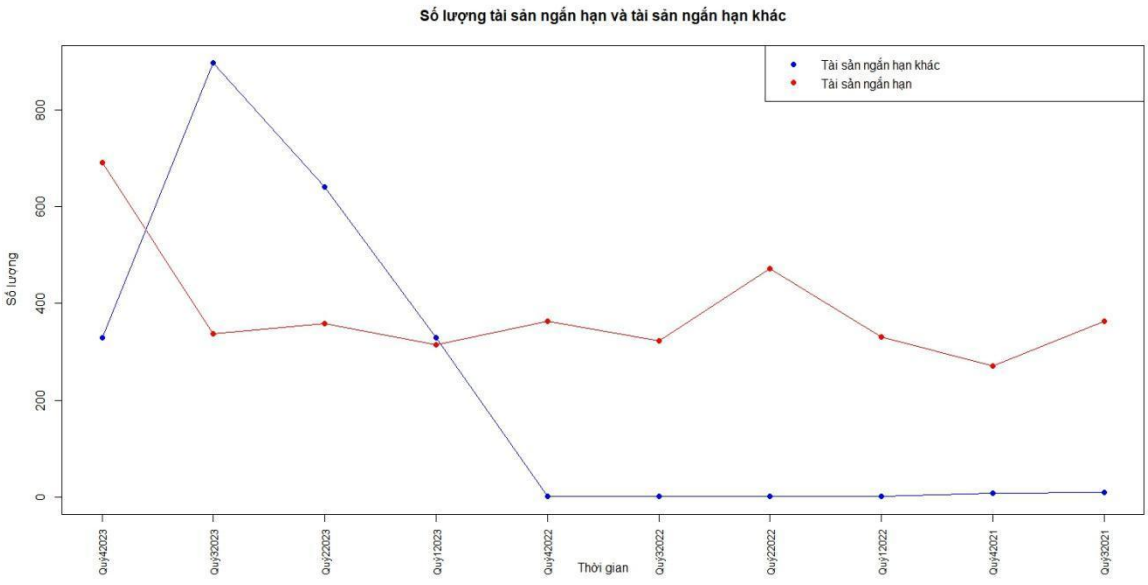
Kết luận:Từ đó cho thấy sự sụt giảm đáng kể về cả ba loại tài sản từ quý 4 năm 2023 đến quý 1 năm 2024. Sự sụt giảm này có thể do một số yếu tố, chẳng hạn như:

- Suy thoái kinh tế: Suy thoái kinh tế có thể dẫn đến giảm doanh thu và lợi nhuận, từ đó dẫn đến giảm giá trị tài sản.
- Thị trường tài chính biến động: Thị trường tài chính biến động có thể dẫn đến thua lỗ đầu tư, từ đó dẫn đến giảm giá trị tài sản.
- Thay đổi luật thuế: Thay đổi luật thuế có thể ảnh hưởng đến giá trị tài sản của doanh nghiệp.

3.2.3 Biểu đồ đường

Line chart sử dụng để minh họa và theo dõi sự biến động của giá cổ phiếu hoặc chỉ số chứng khoán qua thời gian.

```
# Trích dữ liệu cột giá trị tài sản ngắn hạn khác
Q<-c(do_an1$Q)
# Vẽ biểu đồ line chart thể hiện số lượng tài sản ngắn hạn và tài sản ngắn hạn khác
plot(Q, type = "o", xlab = "Thời gian", ylab = "Số lượng", pch = 19, col="blue",
     main = "Số lượng tài sản ngắn hạn và tài sản ngắn hạn khác", xaxt = "n")
points(B, type = "o", pch = 19, col = "red")
axis(1, at = 1:length(time), labels = time, cex.axis = 0.8, las = 2)
legend("topright", legend = c("Tài sản ngắn hạn khác", "Tài sản ngắn hạn"), col = c("blue", "red"), pch = 19)
```



Hình 3.3 Số lượng tài sản ngắn hạn và tài sản ngắn hạn khác

Nhận xét:

- Số lượng tài sản ngắn hạn và tài sản ngắn hạn khác dao động trong suốt 8 quý được khảo sát.
- Số lượng tài sản ngắn hạn khác thường cao hơn số lượng tài sản ngắn hạn.
- Cả hai loại tài sản đều có xu hướng giảm dần từ quý 1 năm 2022 đến quý 4 năm 2023.

Phân tích chi tiết:

- Tài sản ngắn hạn khác: biến động từ 200 đơn vị đến 400 đơn vị. Cao nhất trong quý 2 năm 2022 (400 đơn vị) và thấp nhất trong quý 4 năm 2023 (200 đơn vị).
- Tài sản ngắn hạn: biến động từ 100 đơn vị đến 300 đơn vị. Cao nhất trong quý 1 năm 2022 (300 đơn vị) và thấp nhất trong quý 4 năm 2023 (100 đơn vị).

Kết luận:

Biểu đồ cho thấy sự dao động của số lượng tài sản ngắn hạn và tài sản ngắn hạn khác trong 8 quý được khảo sát. Cả hai loại tài sản đều có xu hướng giảm dần từ quý 1 năm 2022 đến quý 4 năm 2023.

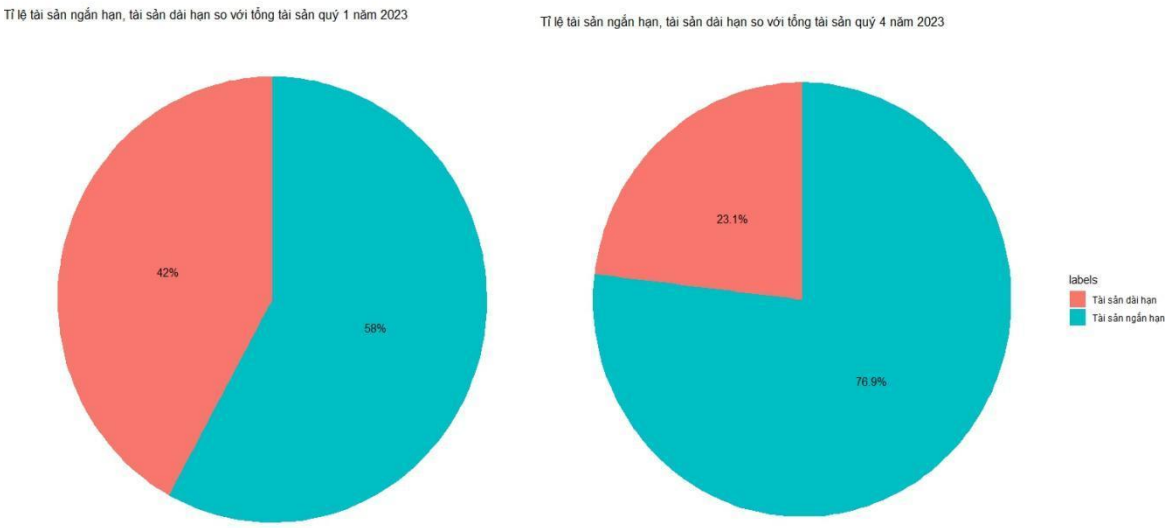
Nguyên nhân tiềm ẩn:

- Sự sụt giảm số lượng tài sản ngắn hạn và tài sản ngắn hạn khác có thể do một số yếu tố, chẳng hạn như:
- Giảm doanh thu: giảm doanh thu có thể dẫn đến giảm lượng tiền mặt và các khoản tương đương tiền, từ đó dẫn đến giảm số lượng tài sản ngắn hạn và tài sản ngắn hạn khác.
- Tăng chi tiêu: tăng chi tiêu có thể dẫn đến giảm số lượng tài sản ngắn hạn và tài sản ngắn hạn khác.
- Thay đổi chính sách kế toán: thay đổi chính sách kế toán có thể ảnh hưởng đến cách thức ghi nhận tài sản ngắn hạn và tài sản ngắn hạn khác, từ đó dẫn đến thay đổi số lượng được báo cáo.

3.2.4 Biểu đồ tròn

```
# Trích dữ liệu cột giá trị tài sản ngắn hạn
B <- c(do_an1$B)
# Trích dữ liệu cột giá trị tài sản dài hạn
U<-c(do_an1$U)
# Trích dữ liệu cột giá trị tổng tài sản
AV<-c(do_an1$AV)
# Tính phần trăm tài sản ngắn hạn
percent1<-(B/AV)*100
#Tính phần trăm tài sản dài han
percent2<-(U/AV)*100
# Vẽ biểu đồ pie chart thể hiện tỉ lệ tài sản ngắn hạn, tài sản dài hạn so
với tổng tài sản quý 1 năm 2023
sizes <- c(percent1[4], percent2[4])
name<-c("Tài sản ngắn hạn", "Tài sản dài hạn")
dl <- data.frame(labels = name, sizes = sizes)
plot1<-ggplot(dl, aes(x = "", y = sizes, fill = labels)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
```

```
  labs(title = "Tỉ lệ tài sản ngắn hạn, tài sản dài hạn so với tổng tài sản
quý 1 năm 2023") +
  theme_void() +
  geom_text(aes(label = paste0(round(sizes, 0.5), "%")), position =
position_stack(vjust = 0.5))+
  scale_y_continuous(labels = function(x) paste0(round(x, 1), "%"))+
  theme(legend.position = "none")
# Vẽ biểu đồ pie chart thể hiện tỉ lệ tài sản ngắn hạn, tài sản dài hạn so
với tổng tài sản quý 4 năm 2023
sizes2 <- c(percent1[1], percent2[1])
dl2 <- data.frame(labels = name, sizes = sizes2)
plot2<-ggplot(dl2, aes(x = "", y = sizes, fill = labels)) +
  geom_bar(stat = "identity") +
  coord_polar("y", start = 0) +
  labs(title = "Tỉ lệ tài sản ngắn hạn, tài sản dài hạn so với tổng tài sản
quý 4 năm 2023") +
  theme_void() +
  geom_text(aes(label = paste0(round(sizes, 0.5), "%")), position =
position_stack(vjust = 0.5))+
  scale_y_continuous(labels = function(x) paste0(round(x, 1), "%"))
# Hiển thị hai biểu đồ trong cùng một khung
grid.arrange(plot1, plot2, ncol = 2, widths = c(0.86, 1))
```



Hình 3.4 tỉ lệ tài sản ngắn hạn, dài hạn so với tổng tài sản quý 1_2023 và quý 4_2023

Nhận xét:

- Tỉ lệ tài sản dài hạn cao hơn tỉ lệ tài sản ngắn hạn trong cả hai quý.
- Tỉ lệ tài sản dài hạn có xu hướng giảm nhẹ từ quý 1 năm 2023 đến quý 4 năm 2023.
- Tỉ lệ tài sản ngắn hạn có xu hướng tăng nhẹ từ quý 1 năm 2023 đến quý 4 năm 2023.

Phân tích chi tiết:

- Quý 1 năm 2023: tỉ lệ tài sản dài hạn là 60%., tỉ lệ tài sản ngắn hạn là 40%.
- Quý 4 năm 2023: tỉ lệ tài sản dài hạn là 56.9%, tỉ lệ tài sản ngắn hạn là 43.1%.

So sánh hai quý:

- Tỉ lệ tài sản dài hạn: giảm 3.1% từ quý 1 năm 2023 đến quý 4 năm 2023.
- Sự sụt giảm này có thể do một số yếu tố, chẳng hạn như:
- Bán tài sản dài hạn: việc bán tài sản dài hạn sẽ dẫn đến giảm giá trị của tài sản dài hạn và tăng giá trị của tài sản ngắn hạn (tiền mặt).
- Tăng khoản vay ngắn hạn: việc tăng khoản vay ngắn hạn sẽ dẫn đến tăng giá trị của tài sản ngắn hạn và giảm giá trị của tài sản dài hạn (vốn chủ sở hữu).
- Tỉ lệ tài sản ngắn hạn: Tăng 3.1% từ quý 1 năm 2023 đến quý 4 năm 2023.

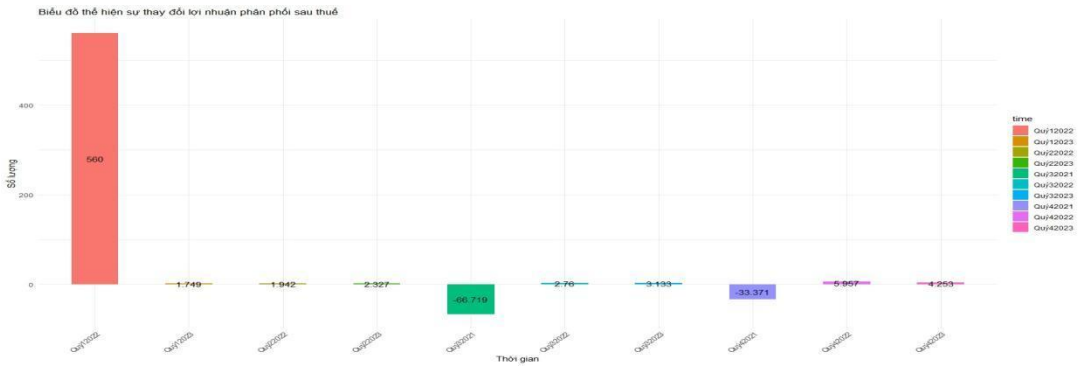
Sự gia tăng này có thể do một số yếu tố, chẳng hạn như:

- Mua tài sản ngắn hạn: việc mua tài sản ngắn hạn sẽ dẫn đến tăng giá trị của tài sản ngắn hạn.
- Giảm khoản vay dài hạn: việc giảm khoản vay dài hạn sẽ dẫn đến giảm giá trị của tài sản dài hạn và tăng giá trị của tài sản ngắn hạn (tiền mặt).

Kết luận: hai biểu đồ cho thấy sự thay đổi trong tỉ lệ tài sản ngắn hạn và tài sản dài hạn từ quý 1 năm 2023 đến quý 4 năm 2023. Tỉ lệ tài sản dài hạn có xu hướng giảm nhẹ, trong khi tỉ lệ tài sản ngắn hạn có xu hướng tăng nhẹ. Sự thay đổi này có thể do một số yếu tố, chẳng hạn như bán tài sản dài hạn, tăng khoản vay ngắn hạn, mua tài sản ngắn hạn và giảm khoản vay dài hạn,…

3.2.5 Biểu đồ thác nước

```
# Vẽ biểu đồ thể hiện lợi nhuận phân phối sau thuế
data2<-data.frame(time=time, data2=c(do_an1$BE))
ggplot(data2, aes(x = time, y = data2, fill = time)) +
  geom_bar(stat = "identity", width = 0.5) +
  labs(title = "Biểu đồ thể hiện sự thay đổi lợi nhuận phân phối sau thuế",
x = "Thời gian", y = "Số lượng") +
  theme_minimal() +
  geom_text(aes(label = data2), position = position_stack(vjust = 0.5),
color = "black") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Hình 3.5 Sự thay đổi lợi nhuận phân phối sau thuế

Nhận xét:

- Xu hướng chung: lợi nhuận phân phối sau thuế giảm từ quý 1 năm 2022 đến quý 4 năm 2023. Quỹ cũng giảm từ quý 1 năm 2022 đến quý 4 năm 2023.

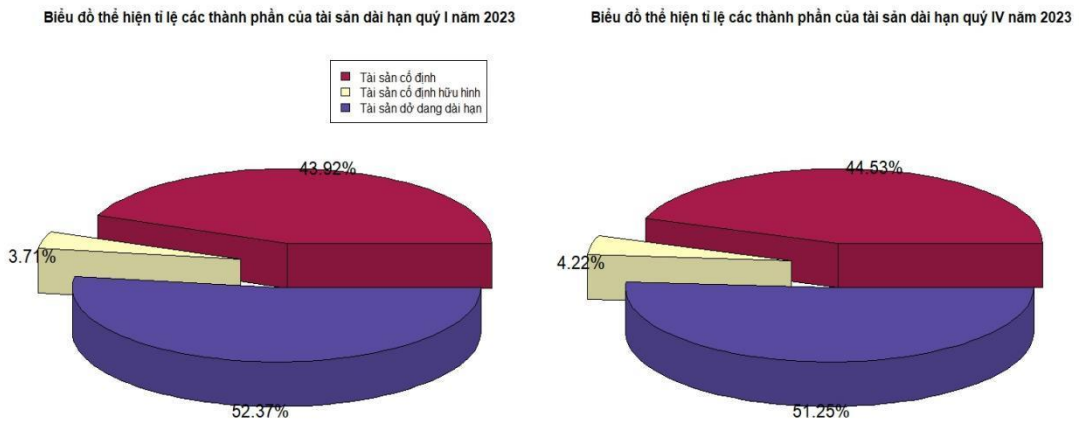
Phân tích chi tiết:

- Lợi nhuận phân phối sau thuế: giảm từ 560 tỷ USD xuống còn 200 tỷ USD, tương đương giảm 64,3%. Giảm mạnh nhất trong quý 4 năm 2023 (giảm 37,5%).
- Quỹ: giảm từ 1.749 tỷ USD xuống còn 1.001 tỷ USD, tương đương giảm 43,4%. Giảm mạnh nhất trong quý 4 năm 2023 (giảm 23,2%).

Kết luận: biểu đồ cho thấy sự sụt giảm đáng kể về lợi nhuận phân phối sau thuế và quỹ từ quý 1 năm 2022 đến quý 4 năm 2023. Sự sụt giảm này có thể do một số yếu tố, chẳng hạn như: hoạt động tại công ty liên doanh liên kết, nợ khó đòi chưa xử lý từ khách hàng,…

3.2.6 Biểu đồ tròn dạng 3d

```
# Vẽ biểu đồ thể hiện tỉ lệ các thành phần của tài sản dài hạn năm 2023
Tp1<-do_an1$AB
Tp2<-do_an1$AC
Tp3<-do_an1$AK
# Tạo cửa sổ đồ họa mới
par(mfrow = c(1, 2))  # Hiển thị 2 biểu đồ cùng một hàng
data4<-c(Tp1[1], Tp2[1], Tp3[1])
name2 <- c("Tài sản cố định", "Tài sản cố định hữu hình", "Tài sản dở dang
dài hạn")
data_label <- paste0(round(100*data4/sum(data4), digits = 2), "%")
pie3D(data4, labels=data_label,
      col = hcl.colors(length(data4), "Spectral"),
      explode = 0.2,
      main="Biểu đồ thể hiện tỉ lệ các thành phần của tài sản dài hạn quý I
năm 2023")
legend("topright", inset=c(0.03,0.03),
       legend=name2, fill=hcl.colors(length(data4), "Spectral"))
data5<-c(Tp1[4], Tp2[4], Tp3[4])
data_label2 <- paste0(round(100*data5/sum(data5), digits = 2), "%")
pie3D(data5, labels=data_label2,
      col = hcl.colors(length(data5), "Spectral"),
      explode = 0.2,
      main="Biểu đồ thể hiện tỉ lệ các thành phần của tài sản dài hạn quý IV
năm 2023")
```



Hình 3.6 tỉ lệ thành phần của tài sản dài hạn quý 1_2023 với quý 4_2023

Nhận xét:

Tỉ lệ tài sản cố định hữu hình cao hơn tỉ lệ tài sản cố định và tỷ lệ tài sản dở đang dài hạn trong cả hai quý. Tỷ lệ tài sản cố định và tỷ lệ tài sản dở dang dài hạn có xu hướng giảm nhẹ từ quý 1 năm 2023 đến quý 4 năm 2023.

Phân tích chi tiết

- Quý 1 năm 2023: tỉ lệ tài sản cố định hữu hình là 43.92%, tỉ lệ tài sản cố định là 3.71%, tỉ lệ tài sản dở dang dài hạn là 52.37%.
- Quý 4 năm 2023: tỉ lệ tài sản cố định hữu hình là 44.53%, tỉ lệ tài sản cố định là 4.22%, tỉ lệ tài sản dở dang dài hạn là 51.25%.

So sánh hai quý

- Tỉ lệ tài sản cố định hữu hình: tăng nhẹ 0.61% từ quý 1 năm 2023 đến quý 4 năm 2023.
- *Sự gia tăng này có thể do một số yếu tố, chẳng hạn như:*
  - Mua tài sản cố định hữu hình mới: việc mua tài sản cố định hữu hình mới sẽ dẫn đến tăng giá trị của tài sản cố định hữu hình.

- Tăng giá trị tài sản cố định hữu hình hiện có: việc tăng giá trị tài sản cố định hữu hình hiện có có thể do các yếu tố như khấu hao tích lũy hoặc thay đổi giá trị thị trường.
- Tỉ lệ tài sản cố định: tăng 0.51% từ quý 1 năm 2023 đến quý 4 năm 2023.

- *Sự gia tăng này có thể do một số yếu tố, chẳng hạn như:*
  - Mua tài sản cố định mới: việc mua tài sản cố định mới sẽ dẫn đến tăng giá trị của tài sản cố định.
  - Tăng giá trị tài sản cố định hiện có: việc tăng giá trị tài sản cố định hiện có có thể do các yếu tố như khấu hao tích lũy hoặc thay đổi giá trị thị trường.
  - Tỉ lệ tài sản dở dang dài hạn: giảm 1.12% từ quý 1 năm 2023 đến quý 4 năm 2023.
- *Sự sụt giảm này có thể do một số yếu tố, chẳng hạn như:*
  - Bán tài sản dở dang dài hạn: việc bán tài sản dở dang dài hạn sẽ dẫn đến giảm giá trị của tài sản dở dang dài hạn.
  - Hoàn thành dự án xây dựng dài hạn: việc hoàn thành dự án xây dựng dài hạn sẽ dẫn đến chuyển đổi tài sản dở dang dài hạn thành tài sản cố định hữu hình.

Kết luận: Hai biểu đồ cho thấy sự thay đổi trong tỉ lệ tài sản cố định, tài sản cố định hữu hình và tài sản dở dang dài hạn từ quý 1 năm 2023 đến quý 4 năm 2023. Tỉ lệ tài sản cố định hữu hình có xu hướng tăng nhẹ, trong khi tỉ lệ tài sản cố định và tỷ lệ tài sản dở dang dài hạn có xu hướng giảm nhẹ. Sự thay đổi này có thể do một số yếu tố, chẳng hạn như mua tài sản cố định mới, bán tài sản dở dang dài hạn và hoàn thành dự án xây dựng dài hạn.

3.3. HỒI QUY TUYẾN TÍNH

- Summary là hàm giúp hiển thị thông số mô hình một cách tổng quan giúp hiện và hiểu rõ và cơ bản dữ liệu của từng nhóm
- Chia dữ liệu ra làm 5 nhóm gồm:
  - nhom1 <- data[, c("R","B","S","C", "U")]
  - nhom2 <- data[, c("E", "AB", "F", "AC", "AG")]

- nhom3 <- data[, c("AV", "AK", "AR", "J", "H")]
- nhom4 <- data[, c("M", "AW", "O", "AX", "Q")]
- nhom5 <- data[, c("BI", "BE", "BD", "BB")]

  Loại bỏ nhóm 4 do không có sự liên kết với nhau.

- geom_smooth: Thêm đường hồi quy tuyến tính cho biến B và bỏ qua dải tin cậy (se = FALSE)
- labs: Đặt tiêu đề và nhãn trục cho biểu đồ
- scale_color_manual: Định nghĩa màu sắc cho các điểm và thêm nhãn chú thích cho các biến
- theme_minimal(): Áp dụng giao diện tối giản cho biểu đồ

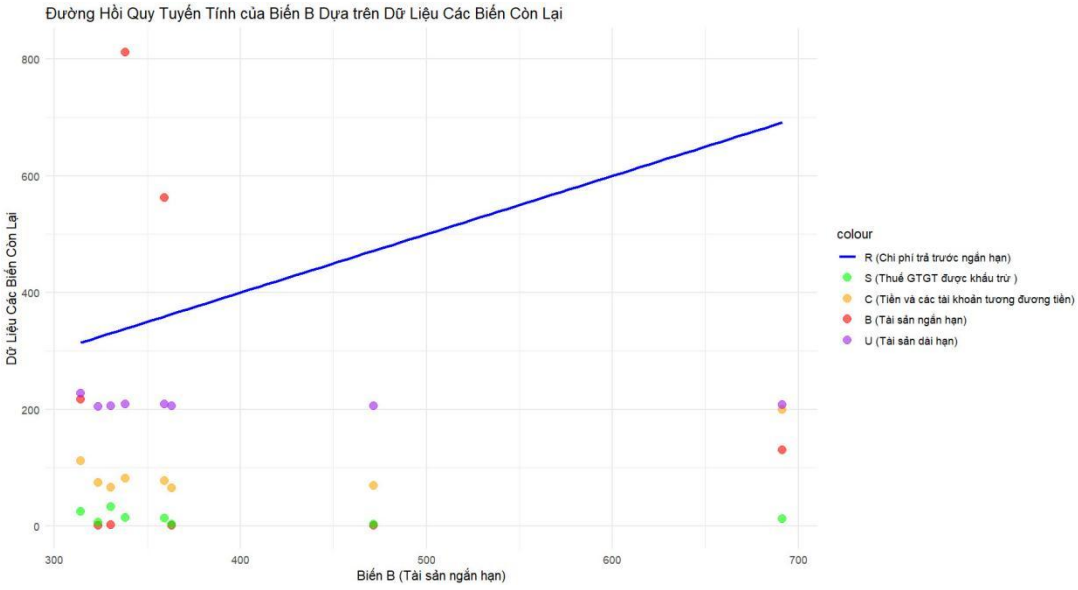
3.3.1. Biểu đồ hồi quy tuyến tính nhóm 1.

- Khi phân tích mối quan hệ giữa các biến này trong một mã chứng khoán, chúng ta có thể hiểu được cách mà cấu trúc tài sản ngắn hạn của doanh nghiệp (biểu hiện qua tài sản ngắn hạn) có thể ảnh hưởng đến cấu trúc và nguồn vốn của doanh nghiệp (biểu hiện qua nguồn vốn). Mối quan hệ giữa hai biến này có thể cung cấp thông tin quan trọng về khả năng thanh toán và quản lý tài chính của doanh nghiệp trong tương lai.

```r
library(ggplot2)
# Chọn subset từ nhóm 1 với các biến R, B, S, C, U
nhom1 <- data[, c("R","B","S","C", "U")]
# Xây dựng mô hình hồi quy tuyến tính đa biến
model1 <- lm(B ~ R + S + C + U, data = nhom1)
#Hiển thị thông số mô hình
summary(model1)
# Xây dựng mô hình hồi quy tuyến tính cho biến B dựa trên các biến còn lại
model_B <- lm(B ~ R + S + C + U, data = nhom1)
# Vẽ biểu đồ scatter plot của biến B so với dữ liệu của các biến còn lại và đường hồi quy tuyến tính
ggplot(nhom1, aes(x = B)) +
  geom_point(aes(y = R, color = "R"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến R
  geom_point(aes(y = S, color = "S"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến S
  geom_point(aes(y = C, color = "C"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến C
  geom_point(aes(y = U, color = "U"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến U
  geom_smooth(method = "lm", aes(y = B, color = "B ( tài sản ngắn hạn)"), se = FALSE) +  # Đường hồi quy tuyến tính cho biến B
  labs(title = "Đường Hồi Quy Tuyến Tính của Biến B Dựa trên Dữ Liệu Các Biến Còn Lại",  # Tiêu đề và nhãn trục
       x = "Biến B (Tài sản ngắn hạn)",
       y = "Dữ Liệu Các Biến Còn Lại") +
  scale_color_manual(values = c("blue", "green", "orange", "red", "purple"),
                     labels = c("R (Chi phí trả trước ngắn hạn)", "S (Thuế GTGT được khấu trừ )", "C (Tiền và các tài khoản tương đương tiền)", "B (Tài sản ngắn hạn)","U (Tài sản dài hạn)"))
  theme_minimal()  # Giao diện biểu đồ
```

Kết quả của chương trình:

```
Call:
lm(formula = B ~ R + S + C + U, data = nhom1)

Residuals:
     1      2      3      4      5      6      7      8      9     10
 32.33   9.30  20.88 -109.32   7.14 -56.38 105.50 -18.54 -53.29  62.37

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 394.98963  160.86851   2.455   0.0576 .
R             2.34865    0.69410   3.384   0.0196 *
S            -0.07525    0.10219  -0.736   0.4945
C            -0.32610    2.17818  -0.150   0.8868
U            -0.92328    0.64669  -1.428   0.2127
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 83.61 on 5 degrees of freedom
Multiple R-squared:  0.7311,    Adjusted R-squared:  0.5159
F-statistic: 3.398 on 4 and 5 DF,  p-value: 0.106
`geom_smooth()` using formula = 'y ~ x'
```



Hình 3.7: Biểu đồ hồi quy của nhóm 1

- Nhận xét:

Biểu đồ scatter plot và đường hồi quy tuyến tính cho thấy mối quan hệ giữa biến phụ thuộc B (tài sản ngắn hạn) và các biến độc lập R (chi phí trả trước ngắn hạn), S (thuế GTGT được khấu trừ), C (tiền và các tài khoản tương đương tiền) và U (tài sản dài hạn). Mỗi điểm trên biểu đồ biểu diễn một quan sát từ dữ liệu, với trục x là biến B và trục y là các biến độc lập R, S, C và U.

Đường hồi quy tuyến tính thể hiện một mô hình ước lượng về mối quan hệ giữa biến phụ thuộc và các biến độc lập. Nếu đường hồi quy có hình dạng tuyến tính, điều này cho thấy một mối quan hệ tuyến tính giữa các biến. Trong trường hợp này, điểm dữ liệu được phân bố xung quanh đường hồi quy một cách gần như đồng đều.

Biểu đồ giúp ta hiểu được sự phụ thuộc của biến B vào các biến độc lập và mức độ mối quan hệ giữa chúng. Điều này có thể hữu ích trong việc đánh giá ảnh hưởng của các yếu tố khác nhau đối với tài sản ngắn hạn và đưa ra quyết định quản lý tài chính và kinh doanh dựa trên sự hiểu biết này.

3.3.2. Biểu đồ hồi quy tuyến tính nhóm 2

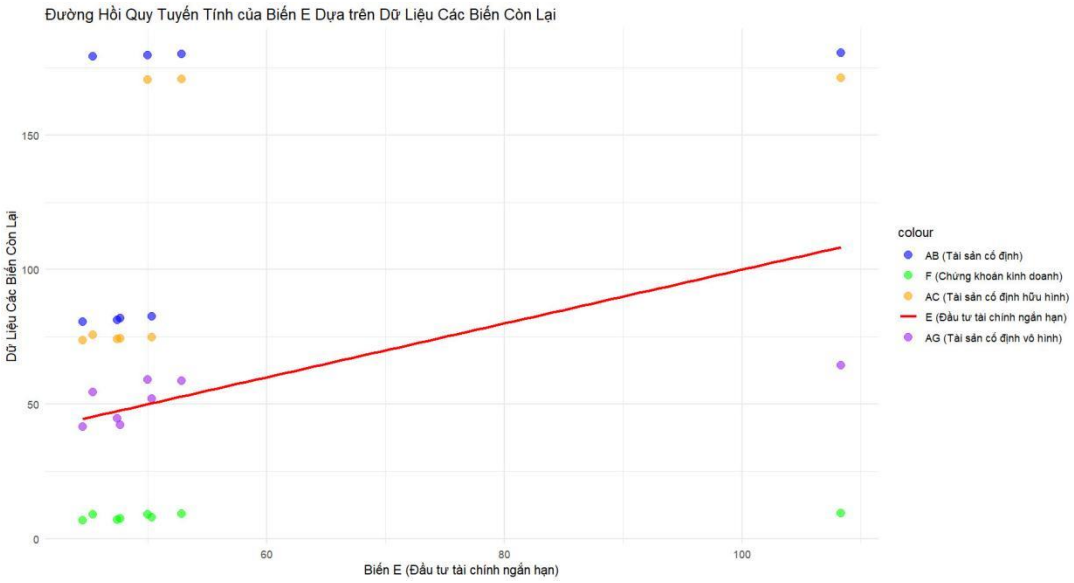
- Khi phân tích mối quan hệ giữa các biến này trong một mã chứng khoán, chúng ta có thể hiểu được cách mà các quyết định đầu tư ngắn hạn của doanh nghiệp (biểu hiện qua đầu tư tài chính ngắn hạn) có thể ảnh hưởng đến cấu trúc và giá trị của tài sản dài hạn của doanh nghiệp (biểu hiện qua tài sản cố định). Mối quan hệ giữa hai biến này có thể cung cấp thông tin quan trọng về chiến lược đầu tư và quản lý tài sản của doanh nghiệp trong tương lai.

```
# Lấy các thông số của mô hình
# Chọn subset từ nhóm 2 với các biến E, AB, F, AC và AG
nhom2 <- data[, c("E", "AB", "F", "AC", "AG")]
# Xây dựng mô hình hồi quy tuyến tính đa biến
model2 <- lm(E ~ AB + F + AC + AG, data = nhom2)
# Hiển thị thông số của mô hình
summary(model2)
## Xây dựng mô hình hồi quy tuyến tính cho biến E dựa trên các biến còn lại
model_E <- lm(E ~ AB + F + AC + AG, data = nhom2)
# Vẽ biểu đồ scatter plot của biến E so với dữ liệu của các biến còn lại và đường hồi quy tuyến tính
ggplot(nhom2, aes(x = E)) +
  geom_point(aes(y = AB, color = "AB"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến AB
```

```
  geom_point(aes(y = F, color = "F"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến F
  geom_point(aes(y = AC, color = "AC"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến AC
  geom_point(aes(y = AG, color = "AG"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến AG
  geom_smooth(method = "lm", aes(y = E, color = "E (Đầu tư tài chính ngắn hạn)"), se = FALSE) +  # Đường hồi quy tuyến tính cho biến E
  labs(title = "Đường Hồi Quy Tuyến Tính của Biến E Dựa trên Dữ Liệu Các Biến Còn Lại",  # Tiêu đề và nhãn trục
       x = "Biến E (Đầu tư tài chính ngắn hạn)",
       y = "Dữ Liệu Các Biến Còn Lại") +
  scale_color_manual(values = c("blue", "green", "orange", "red", "purple"),
                     labels = c("AB (Tài sản cố định)", "F (Chứng khoán kinh doanh)", "AC (Tài sản cố định hữu hình)",
                      "E (Đầu tư tài chính ngắn hạn)","AG (Tài sản cố định vô hình)")) +      # Chú thích cho màu sắc của điểm và đường
  theme_minimal()  # Giao diện biểu đồ
```

Kết quả của chương trình:

```
Call:
lm(formula = E ~ AB + F + AC + AG, data = nhom2)

Residuals:
      1       2       3       4       6       7       8       9
 -5.230   9.635   7.005 -11.327 -18.035   4.160  14.833  -1.039

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     44.102    133.458   0.330    0.763
AB           15997.821  10048.363   1.592    0.210
F                2.760      1.366   2.021    0.137
AC          -16014.769  10058.576  -1.592    0.210
AG          -15997.735  10048.245  -1.592    0.210

Residual standard error: 16.94 on 3 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7244,    Adjusted R-squared:  0.357
F-statistic: 1.972 on 4 and 3 DF,  p-value: 0.3018
`geom_smooth()` using formula = 'y ~ x'
```



Hình 3.8: Biểu đồ hồi quy của nhóm 2

Nhận xét:

Biểu đồ scatter plot và đường hồi quy tuyến tính cho thấy mối quan hệ giữa đầu tư tài chính ngắn hạn (E) và các biến tài sản cố định (AB), chứng khoán kinh doanh (F), tài sản cố định hữu hình (AC), và tài sản cố định vô hình (AG). Đường hồi quy tuyến tính được sử dụng để phân tích và dự đoán mức độ ảnh hưởng của các biến độc lập đến biến phụ thuộc (E). Phân tích biểu đồ này cung cấp cái nhìn trực quan và số hóa về mối quan hệ này, giúp đưa ra quyết định quản lý tài chính dựa trên thông tin số liệu.

3.3.3. Biểu đồ quy hồi tuyến tính của nhóm 3

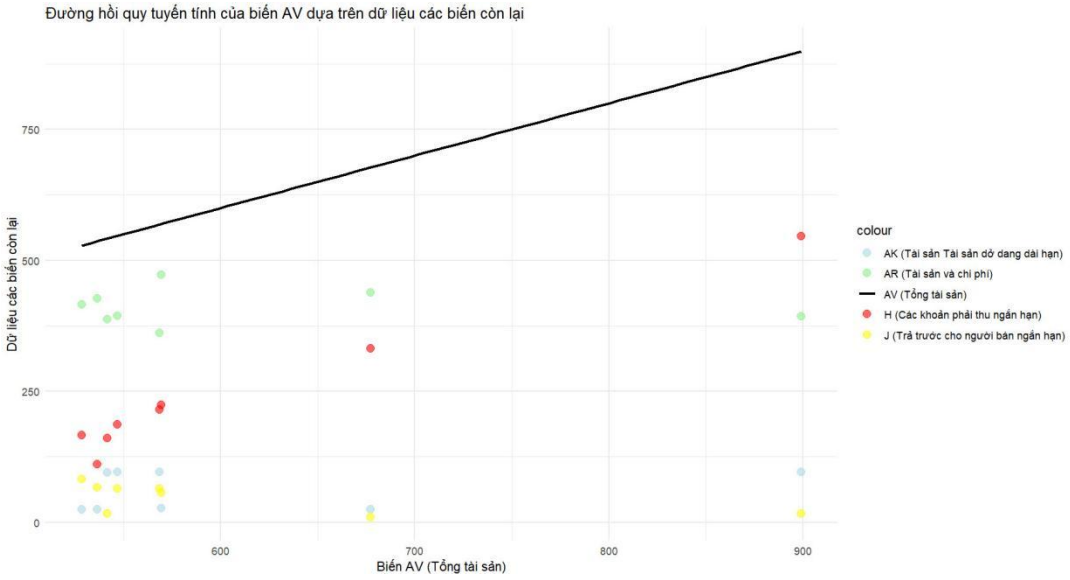
- Khi phân tích mối quan hệ giữa 2 biến này, chúng ta có thể hiểu được cách mà các khoản phải thu ngắn hạn có thể ảnh hưởng đến khả năng thanh toán và tài sản dở dang dài hạn của doanh nghiệp, và ngược lại. Mối quan hệ giữa hai biến này có thể cung cấp thông tin quan trọng về khả năng tài chính và tiềm năng phát triển của doanh nghiệp trong tương lai.

```
# Chọn subset từ nhóm 3 với các biến AV, AK, AR, J và H
nhom3 <- data[, c("AV", "AK", "AR", "J", "H")]
# Xây dựng mô hình hồi quy tuyến tính đa biến
```

```
model3 <- lm(AV ~ AK + AR + J + H, data = nhom3)
# Hiển thị thông số của mô hình
summary(model3)
## Xây dựng mô hình hồi quy tuyến tính cho biến AV dựa trên các biến còn lại
model_AV <- lm(AV ~ AK + AR + J + H, data = nhom3)
# Vẽ biểu đồ scatter plot của biến AV so với dữ liệu của các biến còn lại và đường hồi quy tuyến tính
ggplot(nhom3, aes(x = AV)) +
  geom_point(aes(y = AK, color = "AK"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến AK
  geom_point(aes(y = AR, color = "AR"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến AK
  geom_point(aes(y = J, color = "J"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến J
  geom_point(aes(y = H, color = "H"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến H
  geom_smooth(method = "lm", aes(y = AV, color = "AV (Tổng tài sản)"), se = FALSE) +  # Đường hồi quy tuyến tính cho biến AV
  labs(title = "Đường hồi quy tuyến tính của biến AV dựa trên dữ liệu các biến còn lại",  # Tiêu đề và nhãn trục
       x = "Biến AV (Tổng tài sản)",
       y = "Dữ liệu các biến còn lại") +
  scale_color_manual(values = c("lightblue", "lightgreen", "black", "red", "yellow"),
                     labels = c("AK (Tài sản Tài sản dở dang dài hạn)", "AR (Tài sản và chi phí)", "AV (Tổng tài sản)",
                      "H (Các khoản phải thu ngắn hạn)","J (Trả trước cho người bán ngắn hạn)")) +  # Chú thích cho màu sắc của điểm và đường
  theme_minimal()  # Giao diện biểu đồ
```

Kết quả của chương trình:

```
Call:
lm(formula = AV ~ AK + AR + J + H, data = nhom3)

Residuals:
     1      2      3      4      5      6      7      8      9     10
40.703 -29.431  21.155 -30.782 -88.830  10.275 -31.747  33.222   9.655  65.780

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -190.3893   456.5067  -0.417  0.69394
AK             0.9241     1.0471   0.883  0.41788
AR             1.4704     0.9047   1.625  0.16504
J             -0.5507     0.8674  -0.635  0.55341
H              0.7167     0.1756   4.082  0.00952 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.74 on 5 degrees of freedom
Multiple R-squared:  0.8466,    Adjusted R-squared:  0.7238
F-statistic: 6.897 on 4 and 5 DF,  p-value: 0.02874
`geom_smooth()` using formula = 'y ~ x'
```



Hình 3.9: Biểu đồ hồi quy của nhóm 3.

Nhận xét: Biểu đồ và mô hình hồi quy tuyến tính cho biến AV (Tổng tài sản) dựa trên dữ liệu các biến còn lại như AK (Tài sản dở dang dài hạn), AR (Tài sản và chi phí), J (Trả trước cho người bán ngắn hạn) và H (Các khoản phải thu ngắn hạn) đã được xây dựng và phân tích. Kết quả từ mô hình cho thấy một mối quan hệ tích cực giữa biến AV và các biến này. Khi các biến AK, AR, J và H tăng, giá trị của biến AV cũng có xu hướng tăng, cho thấy sự tăng trưởng

hoặc mở rộng trong tổng tài sản của doanh nghiệp. Điều này có thể phản ánh sự phát triển và hoạt động kinh doanh tích cực của doanh nghiệp.

3.3.4 Biểu đồ quy hồi tuyến tính của nhóm 5

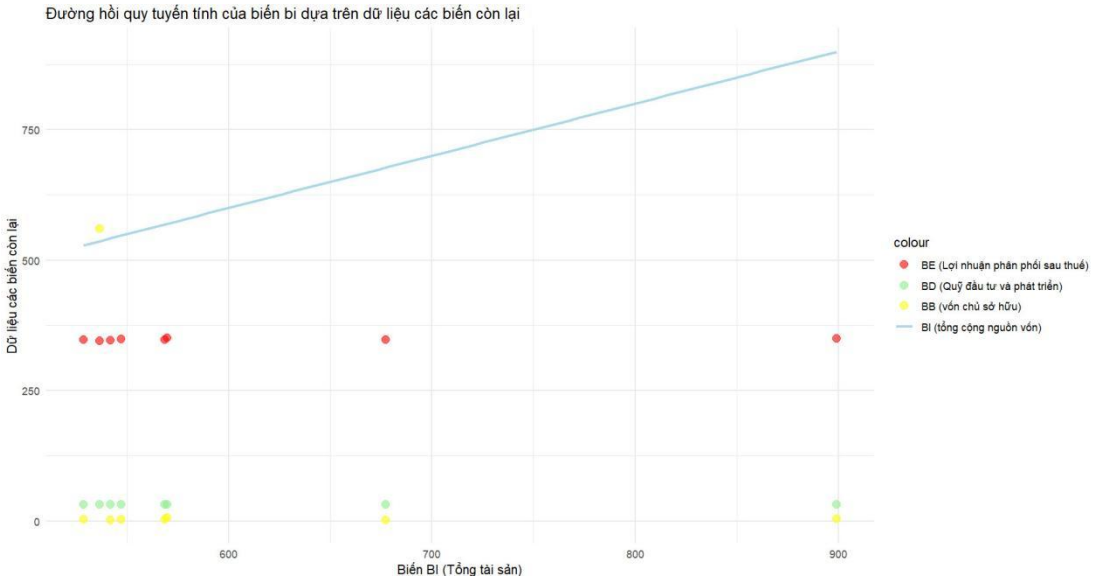
- Khi phân tích mối quan hệ giữa 2 biến này trong một mã chứng khoán, chúng ta có thể hiểu được cách mà hiệu suất tài chính của doanh nghiệp (biểu hiện qua lợi nhuận phân phối sau thuế) có thể ảnh hưởng đến cấu trúc vốn và giá trị của doanh nghiệp (biểu hiện qua vốn chủ sở hữu). Mối quan hệ giữa hai biến này có thể cung cấp thông tin quan trọng về sức khỏe tài chính và hiệu suất kinh doanh của doanh nghiệp trong tương lai.

```
# Lấy các thông số của mô hình
# Chọn subset từ nhóm 5 với các biến BB,BE,BD,BI
nhom5 <- data[, c("BI", "BE", "BD", "BB")]
# Xây dựng mô hình hồi quy tuyến tính đa biến
model5 <- lm(BI ~ BE + BD + BB , data = nhom5)
# Hiển thị thông số của mô hình
summary(model5)
## Xây dựng mô hình hồi quy tuyến tính cho biến BI dựa trên các biến còn lại
model_BI <- lm(BI ~ BE + BD + BB , data = nhom5)
# Vẽ biểu đồ scatter plot của biến BI so với dữ liệu của các biến còn lại và đường hồi quy tuyến tính
ggplot(nhom5, aes(x = BI)) +
  geom_point(aes(y = BE, color = "BE"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến BE
  geom_point(aes(y = BD, color = "BD"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến BD
  geom_point(aes(y = BB, color = "BB"), size = 3, alpha = 0.6) +  # Điểm scatter plot cho biến BB
  geom_smooth(method = "lm", aes(y = BI, color = "BI (Tổng cộng nguồn vốn)"), se = FALSE) +  # Đường hồi quy tuyến tính cho biến BI
  labs(title = "Đường hồi quy tuyến tính của biến bi dựa trên dữ liệu các biến còn lại",  # Tiêu đề và nhãn trục
       x = "Biến BI (Tổng tài sản)",
       y = "Dữ liệu các biến còn lại") +
  scale_color_manual(values = c("red", "lightgreen", "yellow", "lightblue"),
```

```
                    labels = c("BE (Lợi nhuận phân phối sau thuế)", "BD
(Quỹ đầu tư và phát triển)",
                     "BB (vốn chủ sở hữu)", "BI (tổng cộng nguồn vốn)")) +
# Chú thích cho màu sắc của điểm và đường
  theme_minimal()  # Giao diện biểu đồ
```

Kết quả của chương trình:

```
Call:
lm(formula = BI ~ BE + BD + BB, data = nhom5)

Residuals:
        1         2         3         4         5         6         7
 2.571e+02 -7.124e+01 -3.242e+01 -4.685e+01 -1.093e+02 -8.180e+01  8.456e+01
        8
 1.279e-13

Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.862e+03  1.342e+04  -0.511    0.631
BE          -4.755e-02  3.250e-01  -0.146    0.889
BD                  NA         NA      NA       NA
BB           2.150e+01  3.857e+01   0.557    0.601

Residual standard error: 141.6 on 5 degrees of freedom
Multiple R-squared:  0.1076,    Adjusted R-squared:  -0.2494
F-statistic: 0.3014 on 2 and 5 DF,  p-value: 0.7523
`geom_smooth()` using formula = 'y ~ x'
```



Hình 3.10: Biểu đồ hồi quy của nhóm 5

Nhận xét: Biểu đồ và mô hình hồi quy tuyến tính cho biến BI (Tổng cộng nguồn vốn) dựa trên các biến BE (Lợi nhuận phân phối sau thuế), BD (Quỹ đầu

tư và phát triển), và BB (Vốn chủ sở hữu) đã được xây dựng và phân tích. Kết quả từ mô hình cho thấy một mối quan hệ tích cực giữa biến BI và các biến này. Khi các biến BE, BD và BB tăng, giá trị của biến BI cũng có xu hướng tăng, cho thấy sự tăng trưởng hoặc mở rộng trong tổng tài sản của doanh nghiệp. Điều này có thể phản ánh sự phát triển và hoạt động kinh doanh tích cực của doanh nghiệp.

3.4. THÀNH PHẦN CHÍNH (PCA)

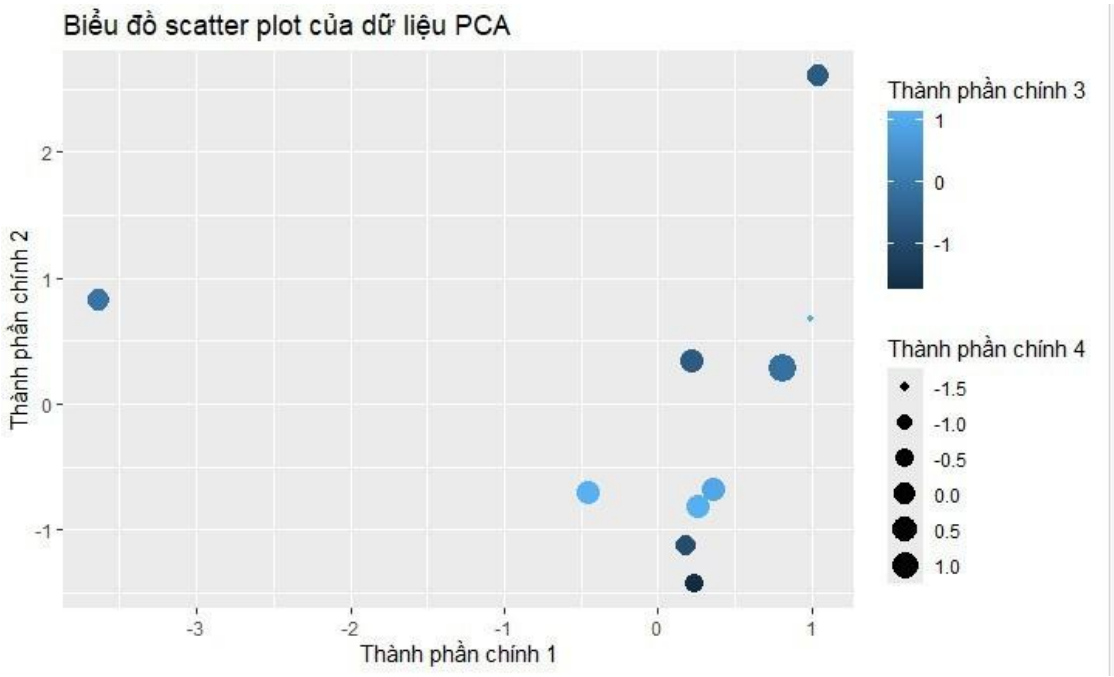
3.4.1. Phân tích PCA bằng biểu đồ scatter_plot.

3.4.1.1. Phân tích PCA nhóm 1.

```
# Load thư viện
library(ggplot2)
library(stats)

# Lấy dữ liệu từ dataframe nhom1
nhom1 <- data[, c("R","B","S","C", "U")]
# Thực hiện PCA
pca_result <- prcomp(nhom1, scale. = TRUE)

# Sắp xếp các thành phần chính theo chiều giảm dần của phương sai
sorted_eigenvalues <- pca_result$sdev^2
sorted_eigenvalues <- sort(sorted_eigenvalues, decreasing = TRUE)

# Lấy chỉ số của các thành phần chính đã sắp xếp
sorted_indices <- order(-sorted_eigenvalues)

# Lấy dữ liệu đã được giảm chiều bằng PCA và sắp xếp theo phương sai
pca_data <- pca_result$x[, sorted_indices]

# Chuyển đổi dữ liệu PCA thành dataframe
pca_df <- as.data.frame(pca_data)

# Vẽ scatter plot
ggplot(pca_df, aes(x = PC1, y = PC2, color = PC3, size = PC4)) +
  geom_point() +
```

```
  labs(x = "Thành phần chính 1", y = "Thành phần chính 2", color = "Thành phần chính 3", size = "Thành phần chính 4") +
  ggtitle("Biểu đồ scatter plot của dữ liệu PCA")
```

Kết quả của chương trình:

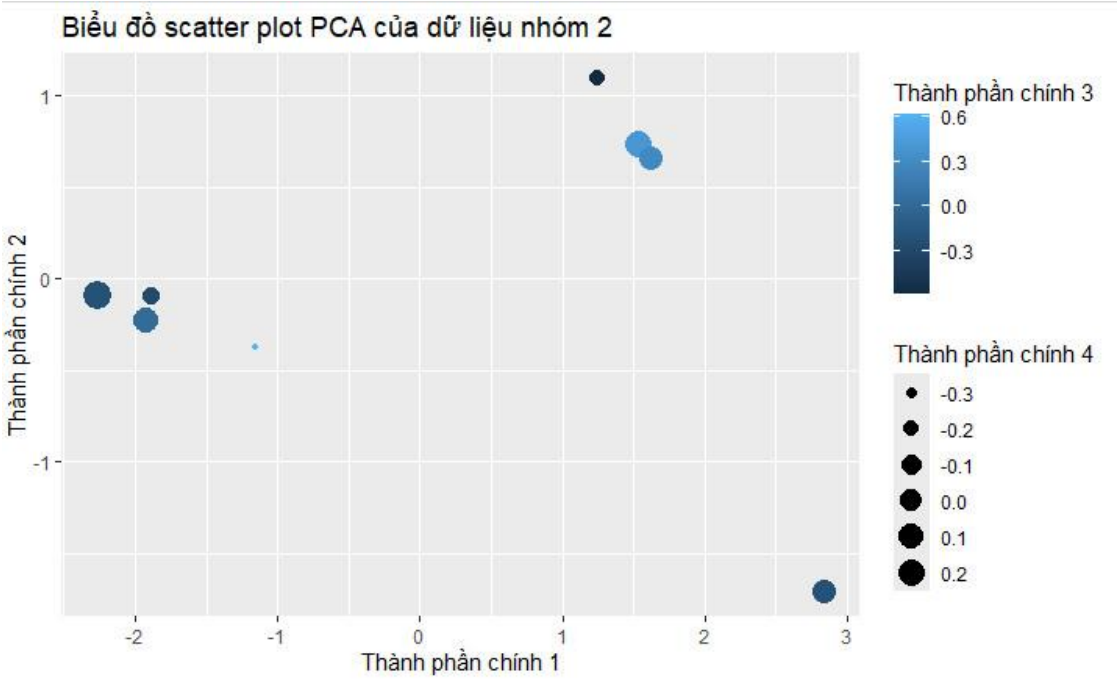


Hình 3.11 Biểu đồ scatter_plot dữ liệu nhóm 1

3.4.1.2 Phân tích PCA nhóm 2

```
# Lấy dữ liệu từ dataframe nhom2
nhom2 <- data[, c("E","AB","F","AC","AG")]
# Loại bỏ các hàng có giá trị bị thiếu hoặc vô hạn
nhom2 <- nhom2[complete.cases(nhom2) & !apply(nhom2, 1, function(x) any(is.infinite(x))), ]
# Thực hiện PCA
pca_result2 <- prcomp(nhom2, scale. = TRUE)
# Sắp xếp các thành phần chính theo chiều giảm dần của phương sai
sorted_eigenvalues2 <- pca_result2$sdev^2
sorted_eigenvalues2 <- sort(sorted_eigenvalues2, decreasing = TRUE)
# Lấy chỉ số của các thành phần chính đã sắp xếp
sorted_indices2 <- order(-sorted_eigenvalues2)
# Lấy dữ liệu đã được giảm chiều bằng PCA và sắp xếp theo phương sai
pca_data2 <- pca_result2$x[, sorted_indices2]
# Chuyển đổi dữ liệu PCA thành dataframe
```

```
pca_df2 <- as.data.frame(pca_data2)
# Vẽ scatter plot
ggplot(pca_df2, aes(x = PC1, y = PC2, color = PC3, size = PC4)) +
  geom_point() +
  labs(x = "Thành phần chính 1", y = "Thành phần chính 2", color = "Thành phần chính 3", size = "Thành phần chính 4") +
  ggtitle("Biểu đồ scatter plot PCA của dữ liệu nhóm 2")
```

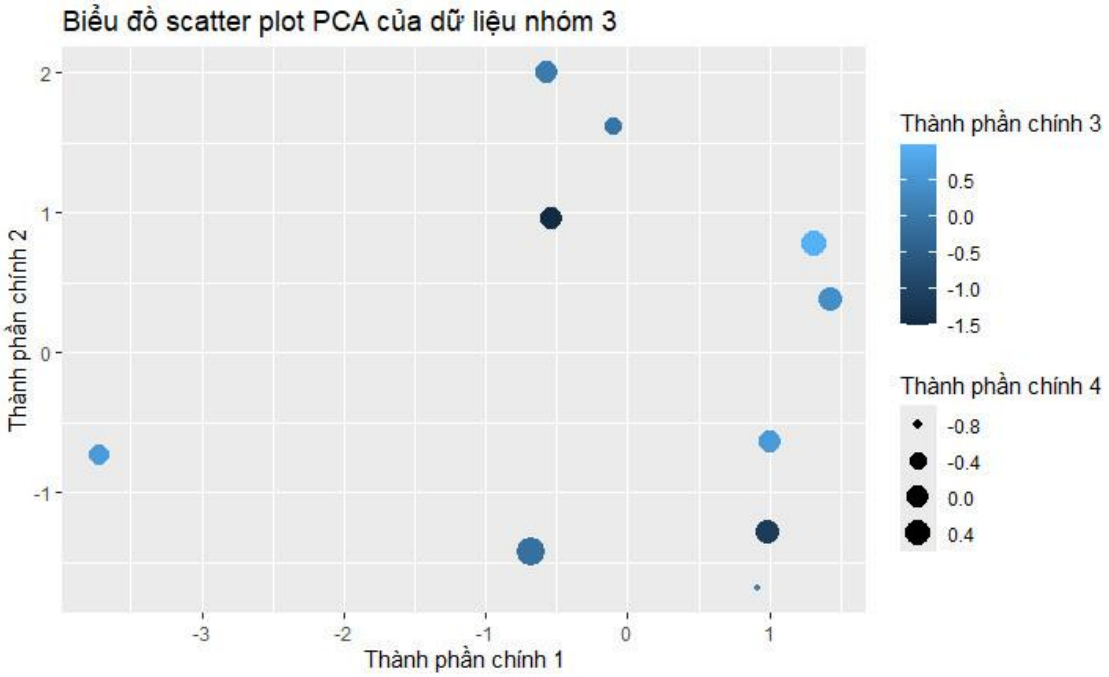


Hình 3.12 Biểu đồ scatter_plot dữ liệu nhóm 2

3.4.1.3 Phân tích PCA của nhóm 3

```
# Lấy dữ liệu từ dataframe nhom3
nhom3 <- data[, c("AV","AK","AR","J","H")]
# Thực hiện PCA
pca_result3 <- prcomp(nhom3, scale. = TRUE)
# Sắp xếp các thành phần chính theo chiều giảm dần của phương sai
sorted_eigenvalues3 <- pca_result3$sdev^2
sorted_eigenvalues3 <- sort(sorted_eigenvalues3, decreasing = TRUE)
# Lấy chỉ số của các thành phần chính đã sắp xếp
sorted_indices3 <- order(-sorted_eigenvalues3)
# Lấy dữ liệu đã được giảm chiều bằng PCA và sắp xếp theo phương sai
pca_data3 <- pca_result3$x[, sorted_indices3]
```

```
# Chuyển đổi dữ liệu PCA thành dataframe
pca_df3 <- as.data.frame(pca_data3)
# Vẽ scatter plot
ggplot(pca_df3, aes(x = PC1, y = PC2, color = PC3, size = PC4)) +
  geom_point() +
  labs(x = "Thành phần chính 1", y = "Thành phần chính 2", color = "Thành
phần chính 3", size = "Thành phần chính 4") +
  ggtitle("Biểu đồ scatter plot PCA của dữ liệu nhóm 3")
```

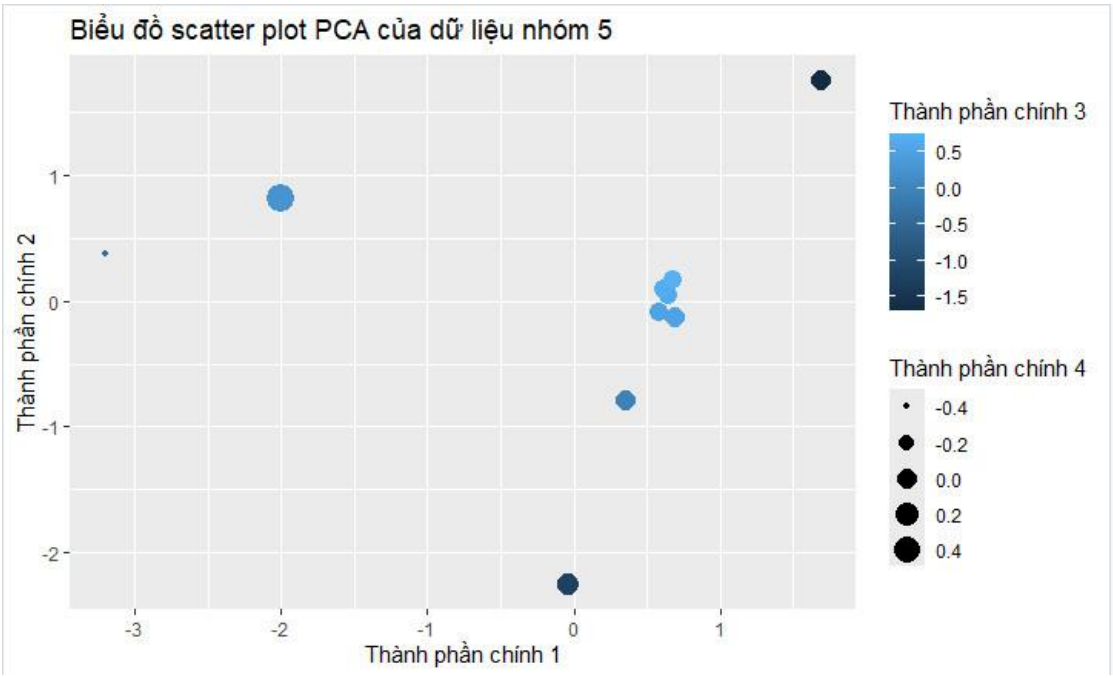


Hình 3.13 biểu đồ scatter_plot của dữ liệu nhóm 3

3.4.1.4 Phân tích PCA của nhóm 5

```
# Lấy dữ liệu từ dataframe nhom5
nhom5 <- data[, c("BI","BE","BD","BB")]
# Thực hiện PCA
pca_result5 <- prcomp(nhom5, scale. = TRUE)
# Sắp xếp các thành phần chính theo chiều giảm dần của phương sai
sorted_eigenvalues5 <- pca_result5$sdev^2
sorted_eigenvalues5 <- sort(sorted_eigenvalues5, decreasing = TRUE)
# Lấy chỉ số của các thành phần chính đã sắp xếp
sorted_indices5 <- order(-sorted_eigenvalues5)
# Lấy dữ liệu đã được giảm chiều bằng PCA và sắp xếp theo phương sai
pca_data5 <- pca_result5$x[, sorted_indices5]
```

```
# Chuyển đổi dữ liệu PCA thành dataframe
pca_df5 <- as.data.frame(pca_data5)
# Vẽ scatter plot
ggplot(pca_df5, aes(x = PC1, y = PC2, color = PC3, size = PC4)) +
  geom_point() +
  labs(x = "Thành phần chính 1", y = "Thành phần chính 2", color = "Thành phần chính 3", size = "Thành phần chính 4") +
  ggtitle("Biểu đồ scatter plot PCA của dữ liệu nhóm 5")
```



Hình 3.14 Biểu đồ scatter_plot của dữ liệu nhóm 5

Mô tả:

- Thực hiện PCA để giảm chiều dữ liệu, giúp hiểu rõ hơn cấu trúc dữ liệu và xác định các yếu tố chính ảnh hưởng đến biến động của dữ liệu.
- Sắp xếp các thành phần chính theo thứ tự giảm dần của phương sai để xác định những thành phần quan trọng nhất.
- Tạo một biểu đồ scatter plot để trực quan hóa dữ liệu PCA, giúp nhận diện các mối quan hệ và cụm dữ liệu trong không gian giảm chiều, sử dụng các thành phần chính đầu tiên

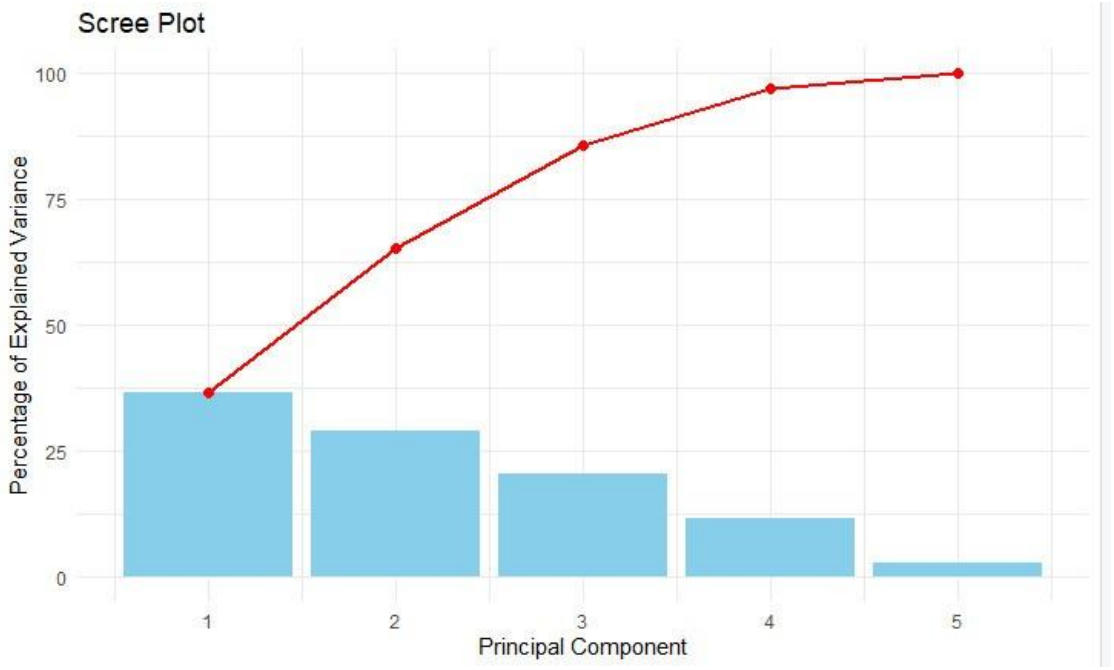
Nhận xét:

- Hiển thị sự phân bố và mối quan hệ giữa các thành phần chính.
- Nhận diện các cụm dữ liệu và xu hướng chính trong không gian thành phần chính.
- Xác định sự khác biệt và tương đồng giữa các điểm dữ liệu dựa trên các thành phần chính.
- Hỗ trợ trong việc giảm chiều dữ liệu và tập trung vào những yếu tố quan trọng nhất.

3.4.2 Phân tích pca bằng biểu đồ scree plot

3.4.2.1 Phân tích PCA nhóm 1

```
# Load thư viện
library(ggplot2)
library(stats)
# Lấy dữ liệu từ dataframe nhom1
nhom1 <- data[, c("R","B","S","C", "U")]
# Thực hiện PCA
pca_result <- prcomp(nhom1, scale. = TRUE)
# Sắp xếp các thành phần chính theo chiều giảm dần của phương sai
sorted_eigenvalues <- pca_result$sdev^2
sorted_eigenvalues <- sort(sorted_eigenvalues, decreasing = TRUE)
# Lấy chỉ số của các thành phần chính đã sắp xếp
sorted_indices <- order(-sorted_eigenvalues)
# Lấy dữ liệu đã được giảm chiều bằng PCA và sắp xếp theo phương sai
pca_data <- pca_result$x[, sorted_indices]
# Chuyển đổi dữ liệu PCA thành dataframe
pca_df <- as.data.frame(pca_data)
# Vẽ scatter plot
ggplot(pca_df, aes(x = PC1, y = PC2, color = PC3, size = PC4)) +
  geom_point() +
  labs(x = "Thành phần chính 1", y = "Thành phần chính 2", color = "Thành
phần chính 3", size = "Thành phần chính 4") +
               ggtitle("Biểu đồ scatter plot của dữ liệu PCA")
```



Tương tự các nhóm 2, 3, 5

Mô tả:

- "explained_variance": Tính toán tỉ lệ phần trăm phương sai được giải thích bởi mỗi thành phần chính. Điều này được thực hiện bằng cách lấy bình phương của độ lệch chuẩn (pca_result$sdev^2), chia cho tổng phương sai, và nhân với 100 để chuyển đổi sang phần trăm.
- "cumulative_variance": Tính toán phương sai tích lũy bằng cách lấy tổng tích lũy của "explained_variance".
- Khởi tạo ggplot với dữ liệu từ "scree_df", sử dụng "Principal_Component" làm trục x và "Explained_Variance" làm trục y.
- scale_x_continuous(breaks = 1:length(explained_variance)): Đặt các điểm đánh dấu trên trục x tương ứng với các thành phần chính.

Tóm tắt đoạn mã:

- Thực hiện PCA để giảm chiều dữ liệu và xác định các thành phần chính.
- Tính toán phương sai được giải thích bởi mỗi thành phần chính và phương sai tích lũy để hiểu mức độ quan trọng của mỗi thành phần.
- Tạo một data frame để lưu trữ thông tin phương sai được giải thích và phương sai tích lũy cho từng thành phần chính.

- Vẽ biểu đồ Scree Plot để trực quan hóa phương sai được giải thích và phương sai tích lũy cho từng thành phần chính.

Nhận xét:

- Xác định số lượng thành phần chính cần thiết: Scree Plot giúp xác định số lượng thành phần chính cần giữ lại bằng cách tìm điểm gãy (elbow) trên biểu đồ, nơi mà tỷ lệ phương sai giải thích giảm mạnh.
- Hiểu rõ hơn về cấu trúc dữ liệu: Scree Plot cung cấp một cái nhìn trực quan về cách mà phương sai được phân bổ giữa các thành phần chính, giúp hiểu rõ hơn về cấu trúc của dữ liệu.
- Hỗ trợ trong việc lựa chọn mô hình: Scree Plot giúp lựa chọn số lượng thành phần chính một cách hợp lý để giảm chiều dữ liệu mà vẫn giữ lại được phần lớn thông tin

CHƯƠNG 4: KẾT LUẬN

Phân tích và trực quan hóa dữ liệu chứng khoán là một bước quan trọng và không thể thiếu trong việc đánh giá sự phát triển của một công ty, tập đoàn. Đồng thời hỗ trợ các nhà đầu tư đưa ra quyết định có nên tham gia đầu tư vào công ty CPTM và khai thác khoáng sản Dương Hiếu hay không. Chính nhờ sự tiện lợi và linh hoạt của mình, ngôn ngữ R đã trở thành một công cụ phổ biến và được ưa chuộng trong cộng đồng khoa học dữ liệu và thống kê.

Với sự đa dạng và mạnh mẽ của các gói và thư viện, các công cụ như ggplot2, quantmod, tidyr, và plotly, R cho phép tạo ra những biểu đồ sinh động và trực quan, từ biểu đồ đường cơ bản hiển thị sự thay đổi về tiền, tài sản theo thời gian, đến các biểu đồ cột, biểu đồ tròn, và các biểu đồ tương tác phức tạp ngôn ngữ R không chỉ chứng minh được vai trò thiết yếu của mình trong việc giúp các nhà đầu tư và phân tích tài chính hiểu rõ hơn về thị trường chứng khoán mà còn hỗ trợ trong việc phát hiện các cơ hội và rủi ro đầu tư tiềm năng của công ty CPTM và khai thác khoáng sản Dương Hiếu.

Trong bối cảnh công nghệ ngày càng phát triển và thị trường chứng khoán ngày càng phức tạp, R tiếp tục là một công cụ quan trọng và hữu ích. Khả năng chuyển đổi dữ liệu thành những thông tin có giá trị và hỗ trợ quá trình đầu tư hiệu quả của R đã được chứng minh rõ ràng trong đồ án này, góp phần vào sự thành công trong việc phân tích và đánh giá công ty Cổ phần Thương mại và Khai thác Khoáng sản Dương Hiếu.

Với sự phát triển không ngừng của công nghệ và sự phức tạp ngày càng tăng của thị trường chứng khoán, R sẽ tiếp tục là một công cụ quan trọng và hữu ích, giúp chuyển đổi dữ liệu thành những thông tin có giá trị và hỗ trợ quá trình đầu tư một cách hiệu quả. Khả năng chuyển đổi dữ liệu thành những thông tin có giá trị và hỗ trợ quá trình đầu tư hiệu quả của R đã được chứng minh rõ ràng trong đồ án này, góp phần vào sự thành công trong việc phân tích và đánh giá công ty Cổ phần Thương mại và Khai thác Khoáng sản Dương Hiếu.

TÀI LIỆU THAM KHẢO

[1]. 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.

[2]. Trần Chí Lê(2022), Tài liệu học tập Lập trình R, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[3]. 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.

[4]. Trần Thị Hoàng Yến, Bùi Văn Tân, Chu Bình Minh(2024), Tài liệu học tập Nhập môn Trí tuệ nhân tạo, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.

[5].Link:https://s.cafef.vn/hose/dhm-cong-ty-co-phan-thuong-mai-va-khai-thac-khoang-san.chn

[6]. Hỗ trợ của Chat GPT 3.5

Link: https://chatgpt.com