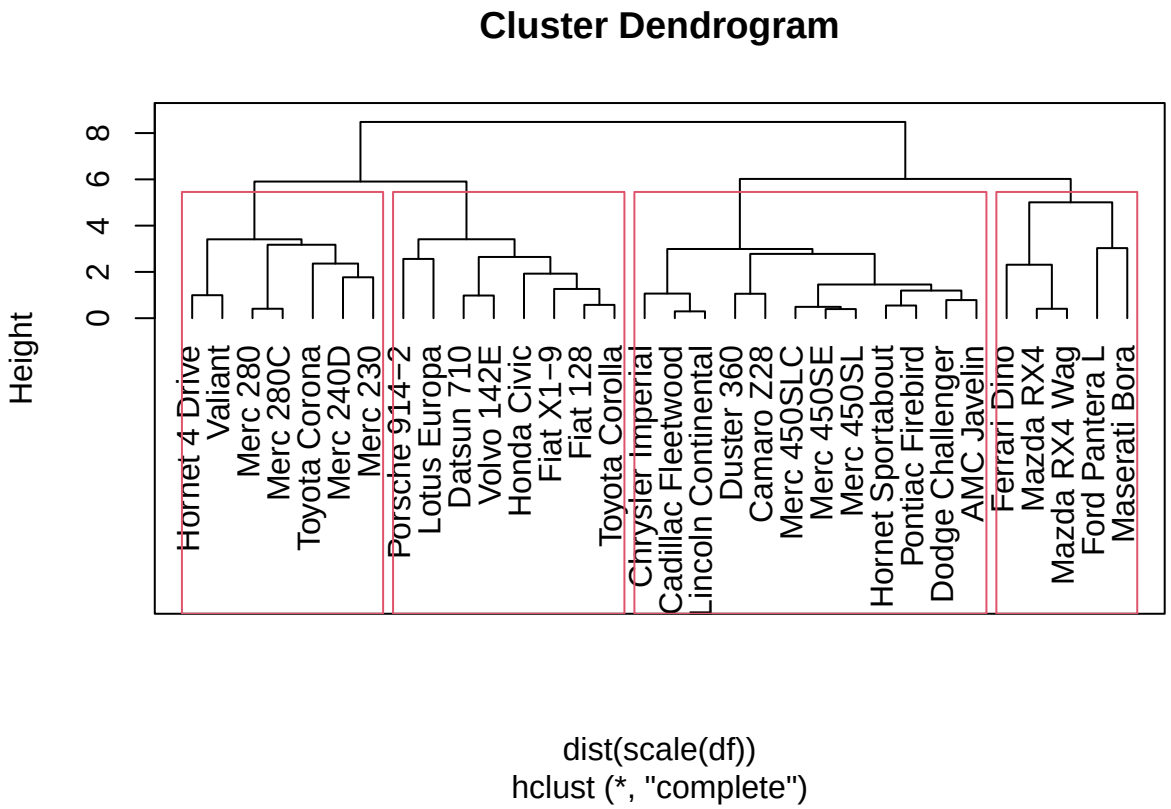
Ćwiczenia 10

2024-01-15

Zadanie 1

```
df <- mtcars
```

```
plot(hclust(dist(scale(df))), frame.plot = TRUE, hang = -1)
rect.hclust(hclust(dist(scale(df))), k = 4)
```

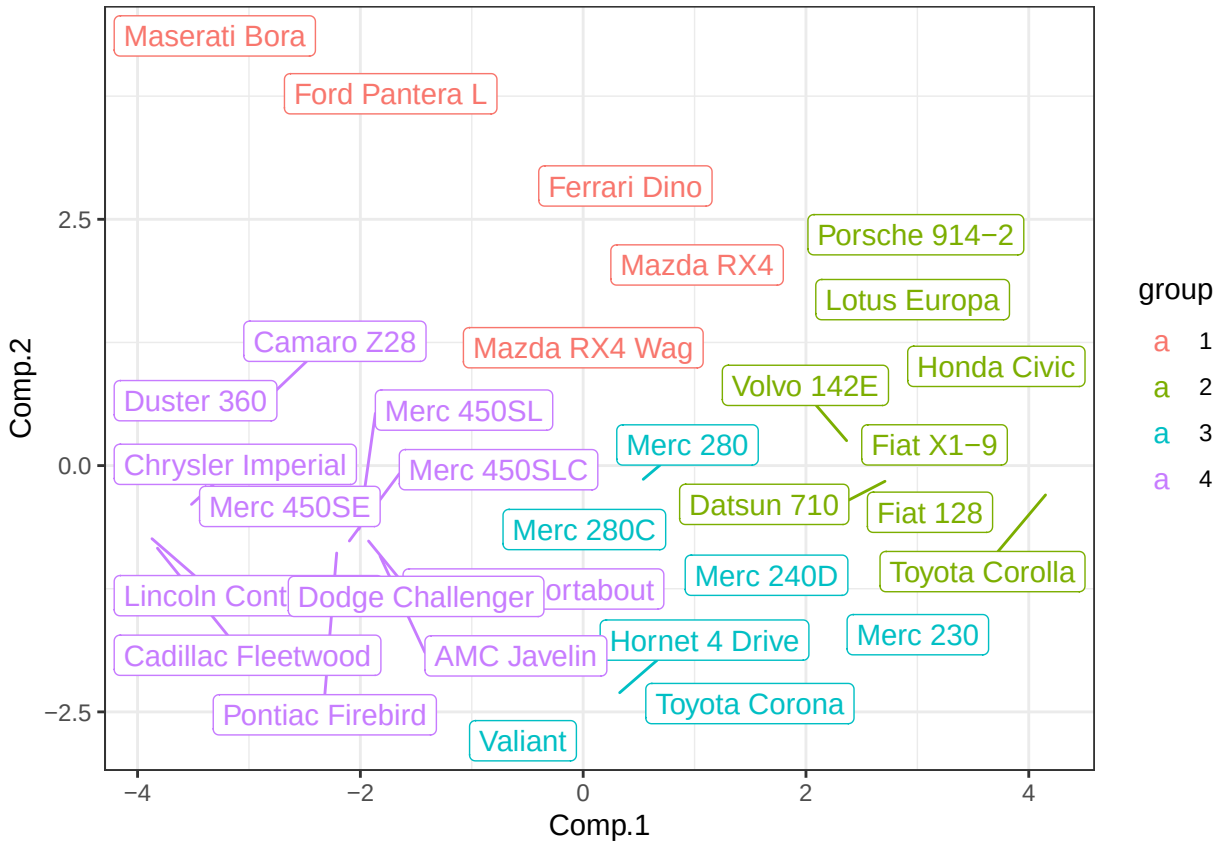



```
cutree(hclust(dist(scale(df)), method = "ward.D2"), k = 4)
```

```
##          Mazda RX4      Mazda RX4 Wag         Datsun 710     Hornet 4 Drive
##                  1                  1                  2                  3
##  Hornet Sportabout          Valiant         Duster 360          Merc 240D
##                  4                  3                  4                  3
##           Merc 230          Merc 280          Merc 280C         Merc 450SE
##                  3                  3                  3                  4
##         Merc 450SL        Merc 450SLC  Cadillac Fleetwood Lincoln Continental
##                  4                  4                  4                  4
##  Chrysler Imperial          Fiat 128         Honda Civic      Toyota Corolla
##                  4                  2                  2                  2
##       Toyota Corona    Dodge Challenger         AMC Javelin          Camaro Z28
```

```
##                   3                   4                   4                   4
##    Pontiac Firebird            Fiat X1-9       Porsche 914-2        Lotus Europa
##                   4                   2                   2                   2
##      Ford Pantera L        Ferrari Dino       Maserati Bora          Volvo 142E
##                   1                   1                   1                   2
```

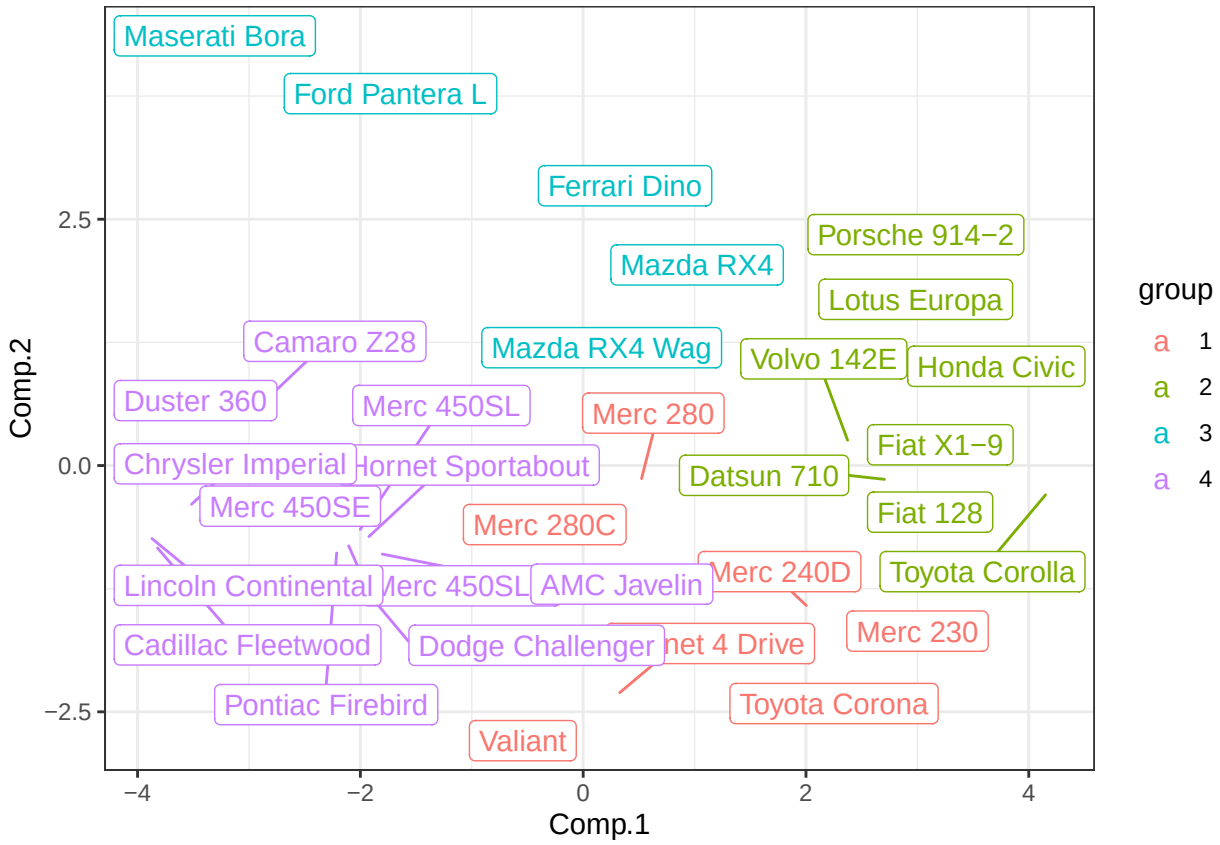
```
df |>
  scale() |>
  princomp() |>
  predict() |>
  as_tibble() |>
  mutate(group = cutree(hclust(dist(scale(df)), method = "ward.D2"), k = 4) |> factor(),
         name = rownames(mtcars)) |>
  ggplot(aes(col = group, y = Comp.2, x = Comp.1, label = name)) +
  ggrepel::geom_label_repel(max.overlaps = 17) +
  theme_bw()
```




```
cl <- kmeans(scale(df), centers = 4)
```

```
df |>
  scale() |>
  princomp() |>
  predict() |>
  as_tibble() |>
  mutate(group = cl$cluster |> factor(),
         name = rownames(mtcars)) |>
  ggplot(aes(col = group, y = Comp.2, x = Comp.1, label = name)) +
  ggrepel::geom_label_repel(max.overlaps = 17) +
```

```
  theme_bw()
```

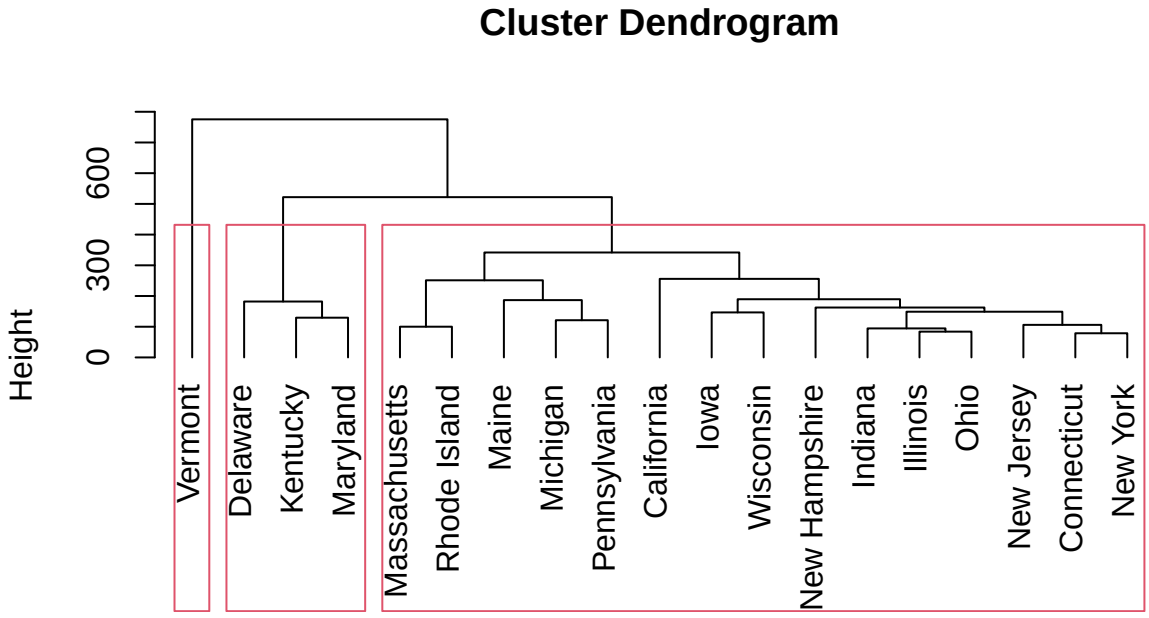


Zadanie 2

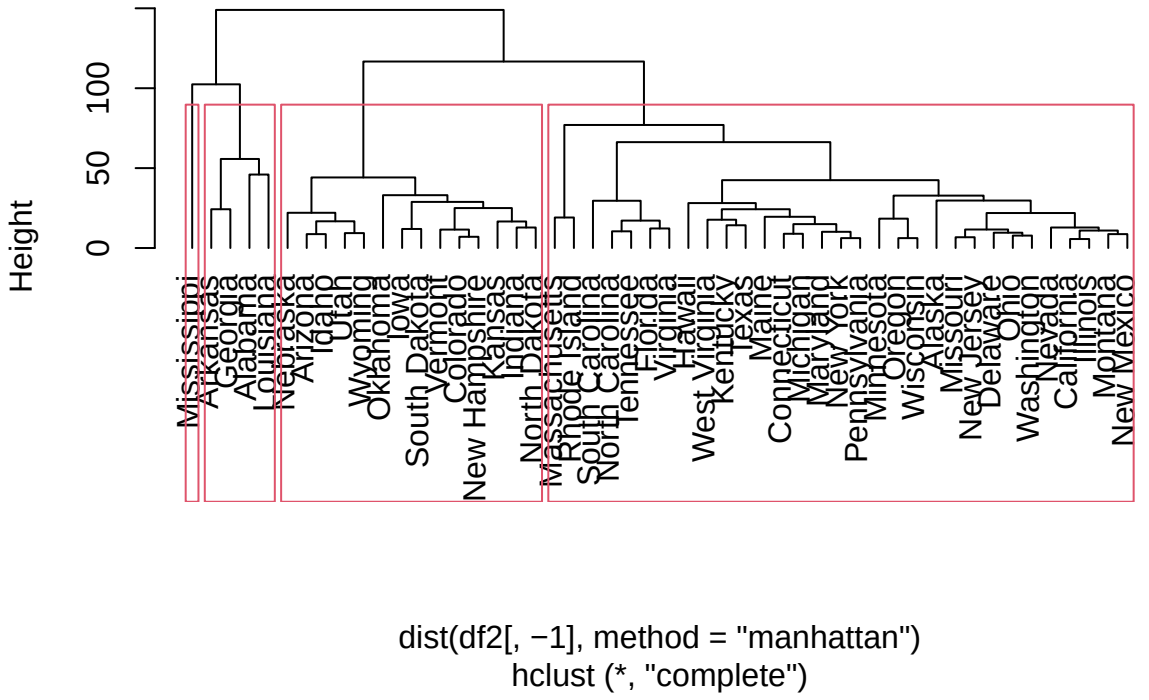
```
df1 <- cluster::votes.repub |>
  na.omit() |>
  as_tibble(rownames = "State")

df2 <- cluster::votes.repub |>
  t() |>
  na.omit() |>
  t() |>
  as_tibble(rownames = "State")
```

```
model_hlucst1 <- hclust(dist(df1[, -1], method = "manhattan"),
                        method = "complete")
model_hlucst2 <- hclust(dist(df2[, -1], method = "manhattan"),
                        method = "complete")
```

```
plot(model_hlucst1, labels = df1$State, hang = -1)
rect.hclust(model_hlucst1, k = 3)
```

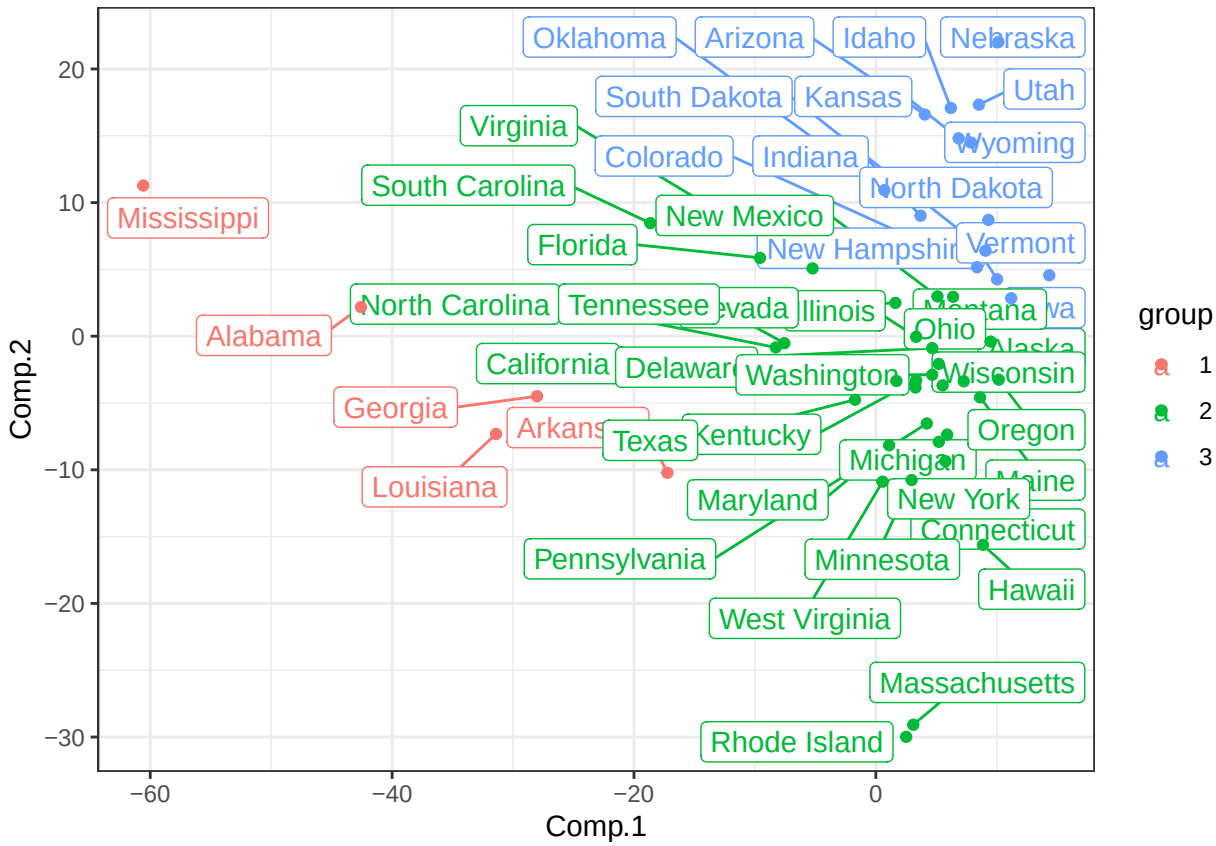



```
plot(model_hlucst2, labels = df2$State, hang = -1)
rect.hclust(model_hlucst2, k = 4)
```

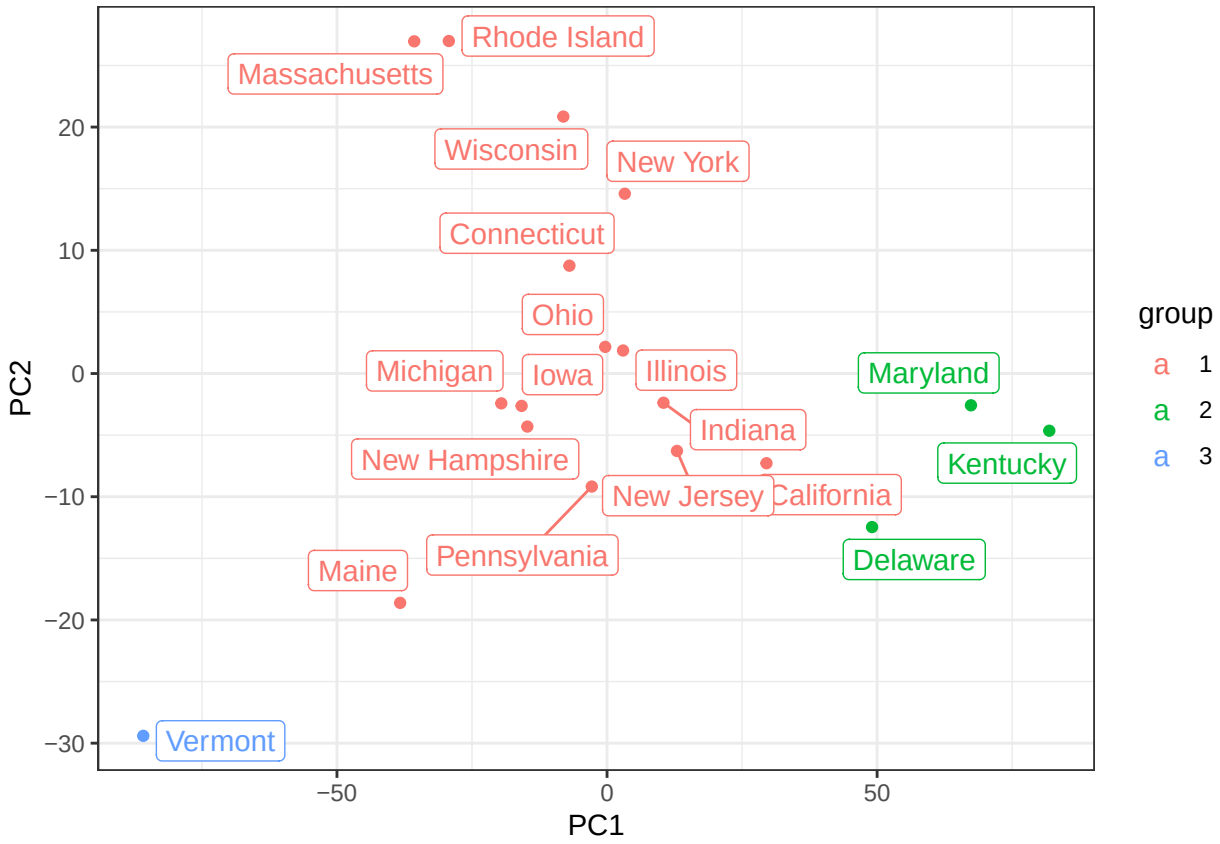


```
df2 |>
  select(-State)|>
  princomp() |>
  predict() |>
  as_tibble() |>
  transform(State = df2$State,
            group = cutree(model_hlucst2, k = 3) |> factor()) |>
  ggplot(aes(col = group, y = Comp.2, x = Comp.1, label = State)) +
  ggrepel::geom_label_repel(max.overlaps = 26) +
  geom_point() +
  theme_bw()
```

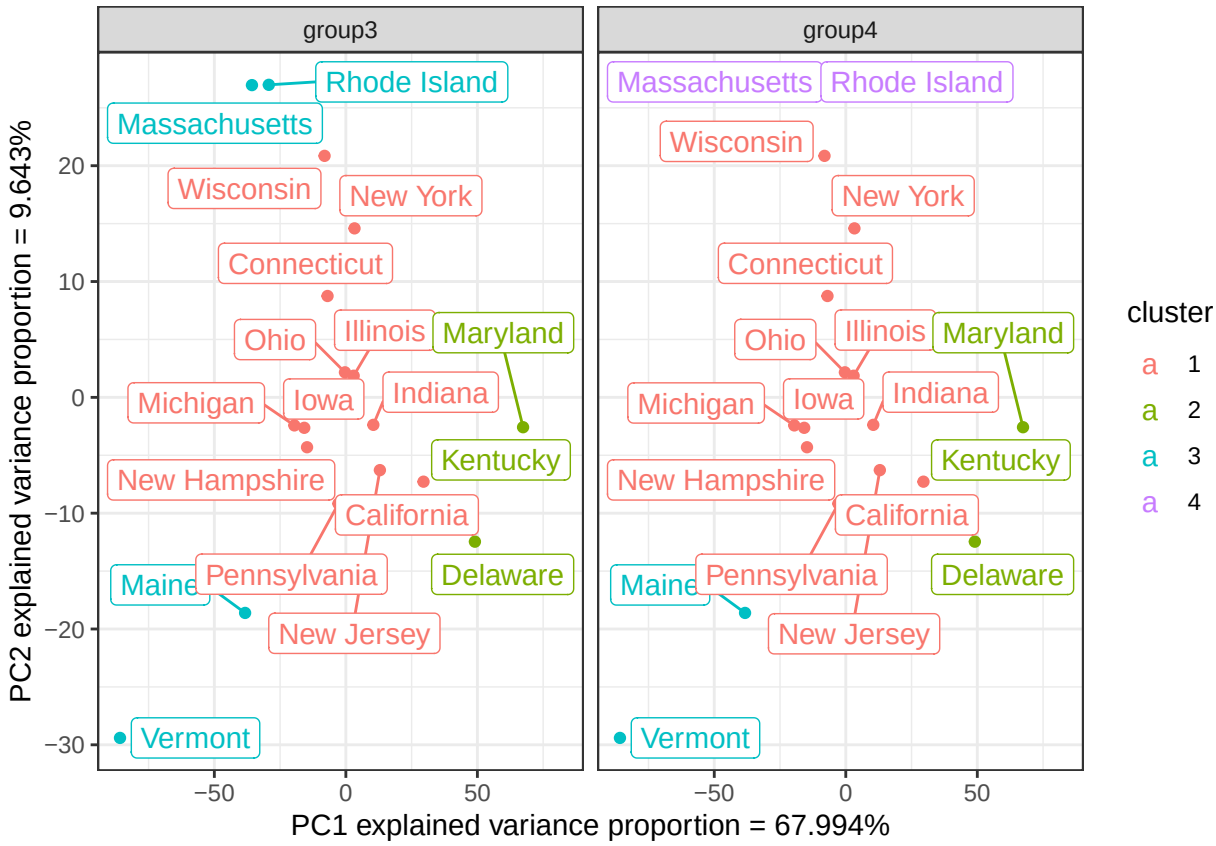
```
## Warning: ggrepel: 2 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



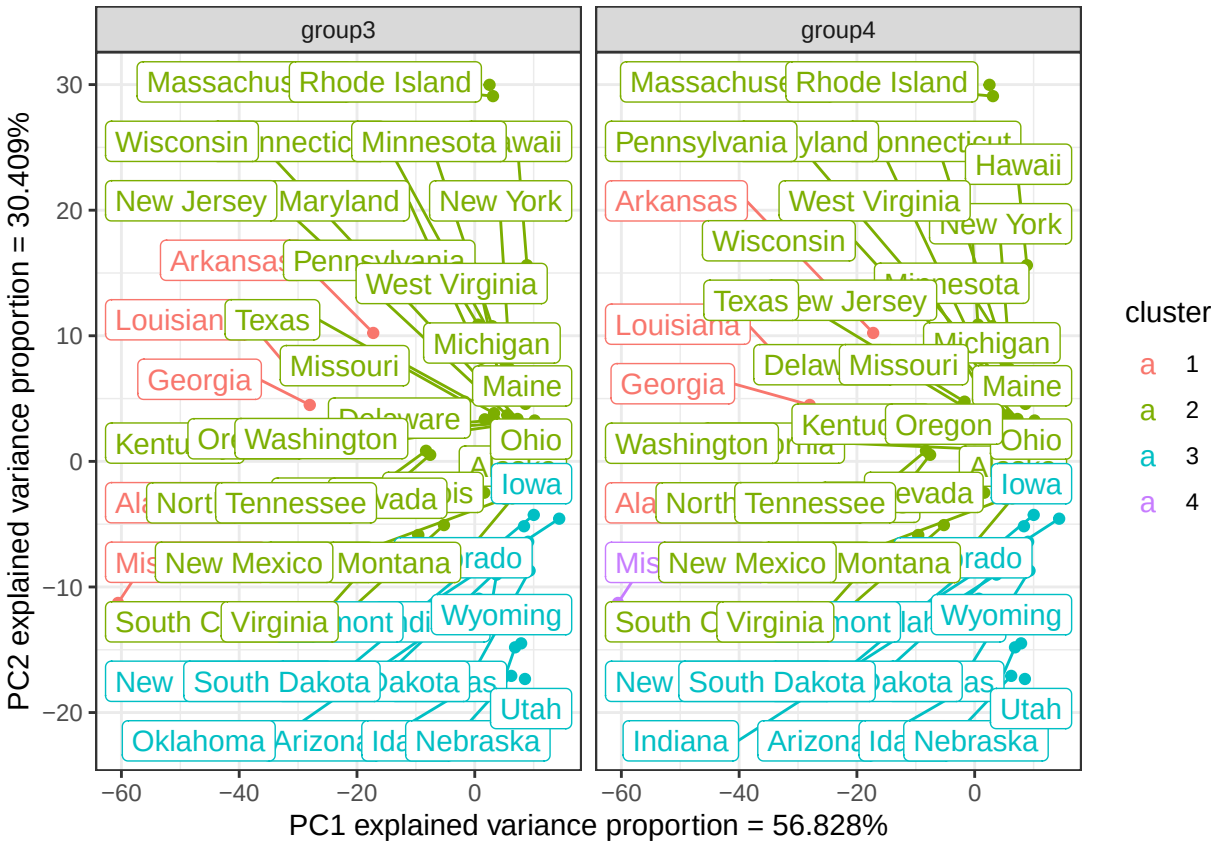

```
df1 |>
  select(-State)|>
  prcomp() |>
  predict() |>
  as_tibble() |>
  transform(State = df1$State,
            group = cutree(model_hlucst1, k = 3) |> factor()) |>
  ggplot(aes(col = group, y = PC2, x = PC1, label = State)) +
  geom_point() +
  ggrepel::geom_label_repel(max.overlaps = 10) +
  theme_bw()
```




```
df1 |>
  select(-State)|>
  prcomp() -> xx

xx |>
  predict() |>
  as_tibble() |>
  transform(State  = df1$State,
            group3 = cutree(hclust(dist(df1[, -1], method = "manhattan"),
                       method = "ward.D"), k = 3) |> factor(),
            group4 = cutree(hclust(dist(df1[, -1], method = "manhattan"),
                       method = "ward.D"), k = 4) |> factor()) |>
  pivot_longer(cols = c("group3", "group4"),
               values_to = "cluster",
               names_to = "number_of_clusters") |>
  ggplot(aes(y = PC2, x = PC1, label = State, color = cluster)) +
  geom_point() +
  facet_grid(cols = vars(number_of_clusters)) +
  ggrepel::geom_label_repel(max.overlaps = 20) +
  theme_bw() +
  xlab(paste0("PC1 explained variance proportion = ",
              summary(xx)$importance[2, 1] * 100,
              "%")) +
  ylab(paste0("PC2 explained variance proportion = ",
              summary(xx)$importance[2, 2] * 100,
              "%"))
```




```
df2 |>
  select(-State)|>
  prcomp() -> xx2

xx2 |>
  predict() |>
  as_tibble() |>
  transform(State  = df2$State,
            group3 = cutree(model_hlucst2, k = 3) |> factor(),
            group4 = cutree(model_hlucst2, k = 4) |> factor()) |>
  pivot_longer(cols = c("group3", "group4"),
               values_to = "cluster",
               names_to = "number_of_clusters") |>
  ggplot(aes(y = PC2, x = PC1, label = State, color = cluster)) +
  geom_point() +
  facet_grid(cols = vars(number_of_clusters)) +
  ggrepel::geom_label_repel(max.overlaps = 35) +
  theme_bw() +
  xlab(paste0("PC1 explained variance proportion = ",
              summary(xx2)$importance[2, 1] * 100,
              "%")) +
  ylab(paste0("PC2 explained variance proportion = ",
              summary(xx2)$importance[2, 2] * 100,
              "%"))
```


group3
group4
Massachus
Rhode Island
Wisconsin
Minnesota
New Jersey
Maryland
New York
Arkansas
West Virginia
Louisian
Texas
Michigan
Missouri
Georgia
Maine
Ohio
Washington
Iowa
Tennessee
New Mexico
Montana
Wyoming
South C
Virginia
New
South Dakota
Utah
Oklahoma
Arizona
Nebraska
Massachuse
Pennsylvania
Hawaii
Arkansas
West Virginia
New York
Wisconsin
Texas
New Jersey
Louisiana
Michigan
Missouri
Georgia
Maine
Oregon
Ohio
Washington
Iowa
Tennessee
New Mexico
Montana
Wyoming
South C
Virginia
New
South Dakota
Utah
Indiana
Arizona
Nebraska
30
20
10
0
−10
−20
−60
−40
−20
0
PC2 explained variance proportion = 30.409%
PC1 explained variance proportion = 56.828%
cluster
a 1
a 2
a 3
a 4