DATA VISUALIZATION

Part 2: Forest plot & Map

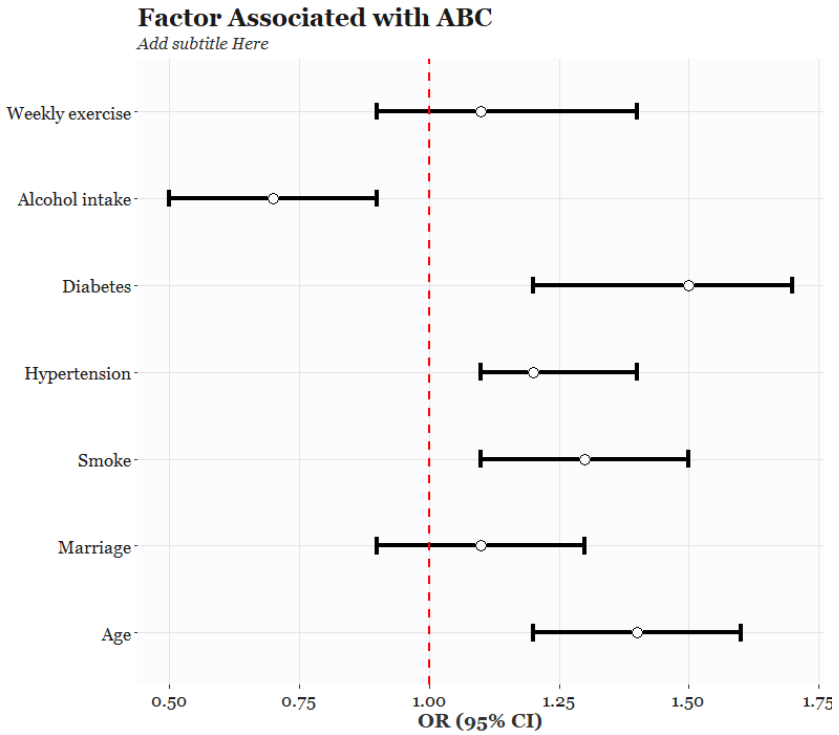
Khuong Quynh Long
Ha Noi, 03/2019
https://gitlab.com/LongKhuong/data-visualization

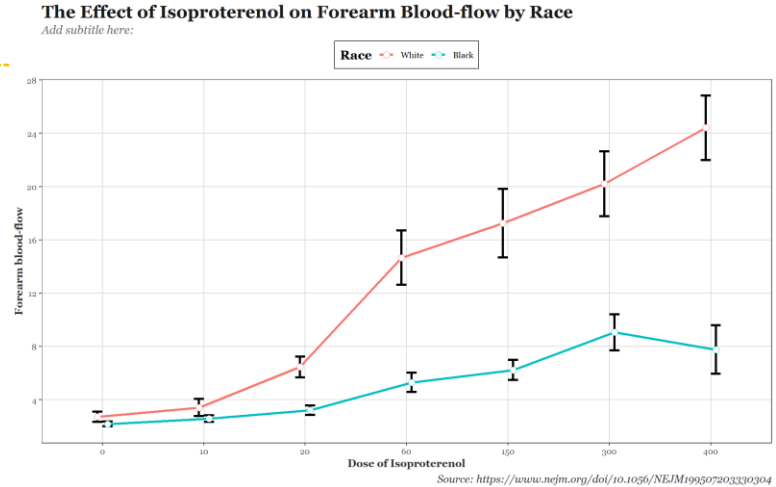
Nội dung

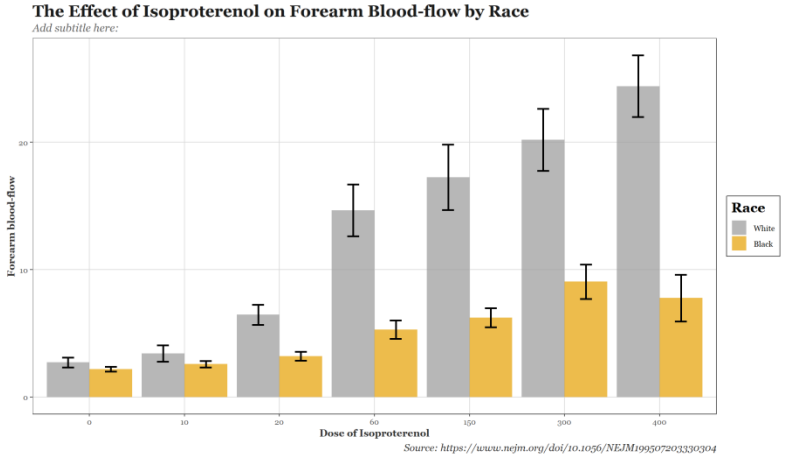
- Cách vẽ Errorbar
- Cách extract kết quả sang dataframe ➔ Forest plot
- Lấy Map từ Package “raster”
- Merge dữ liệu với shape file ➔ Map
- Tùy chỉnh Map theo các marker

Forest plot

Errorbar “family”







Dữ liệu

```
forestPlot <- tribble(
~factor, ~OR, ~low, ~high,
#------/------/------/------/
"Age", 1.4,1.2,1.6,
"Marriage", 1.1, 0.9, 1.3,
"Smoke", 1.3, 1.1, 1.5,
"Hypertension", 1.2, 1.1, 1.4,
"Diabetes",1.5, 1.2, 1.7,
"Alcohol intake",0.7,0.5,0.9,
"Weekly exercise",1.1, 0.9, 1.4 )

forestPlot
```

Dữ liệu

```
## # A tibble: 7 x 4
##   factor            OR   low  high
##   <chr>          <dbl> <dbl> <dbl>
## 1 Age              1.4   1.2   1.6
## 2 Marriage         1.1   0.9   1.3
## 3 Smoke            1.3   1.1   1.5
## 4 Hypertension     1.2   1.1   1.4
## 5 Diabetes         1.5   1.2   1.7
## 6 Alcohol intake   0.7   0.5   0.9
## 7 Weekly exercise  1.1   0.9   1.4
```

Forest plot

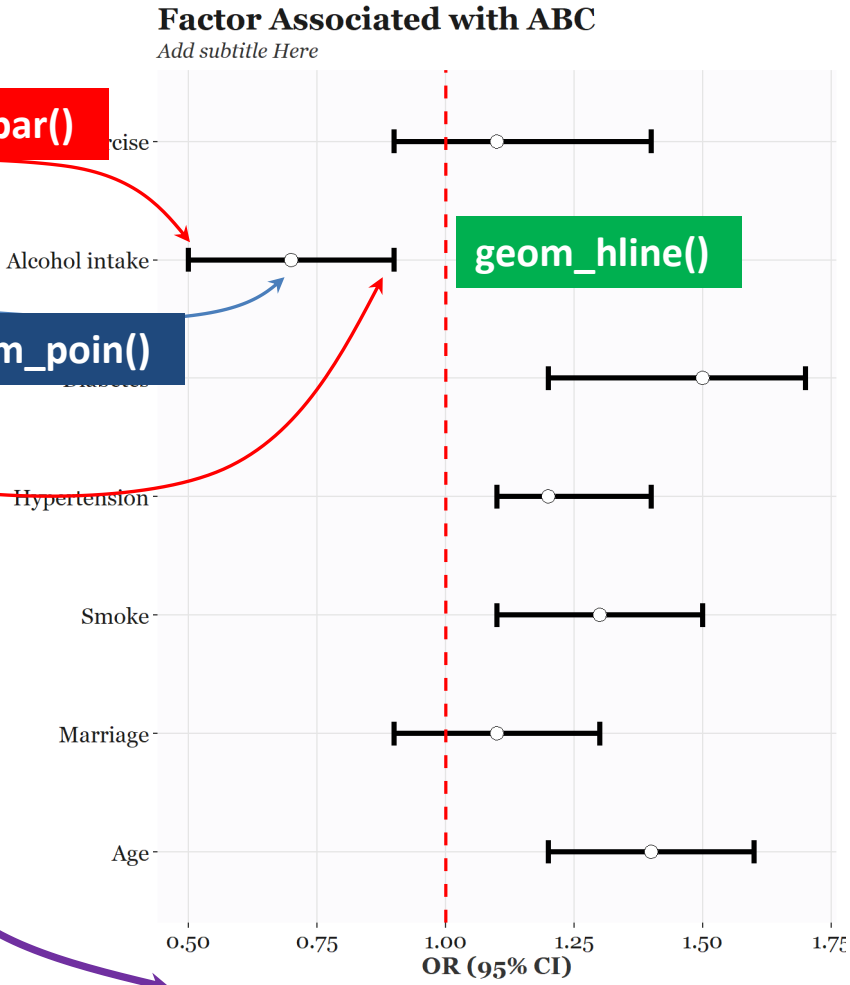
```
## # A tibble: 7 x 4
##   factor            OR   low  high
##   <chr>          <dbl> <dbl> <dbl>
## 1 Age              1.4   1.2   1.6
## 2 Marriage         1.1   0.9   1.3
## 3 Smoke            1.3   1.1   1.5
## 4 Hypertension     1.2   1.1   1.4
## 5 Diabetes         1.5   1.2   1.7
## 6 Alcohol intake   0.7   0.5   0.9
## 7 Weekly exercise  1.1   0.9   1.4
```

Layer: geom_errorbar()

Layer: geom_poin()

geom_hline()

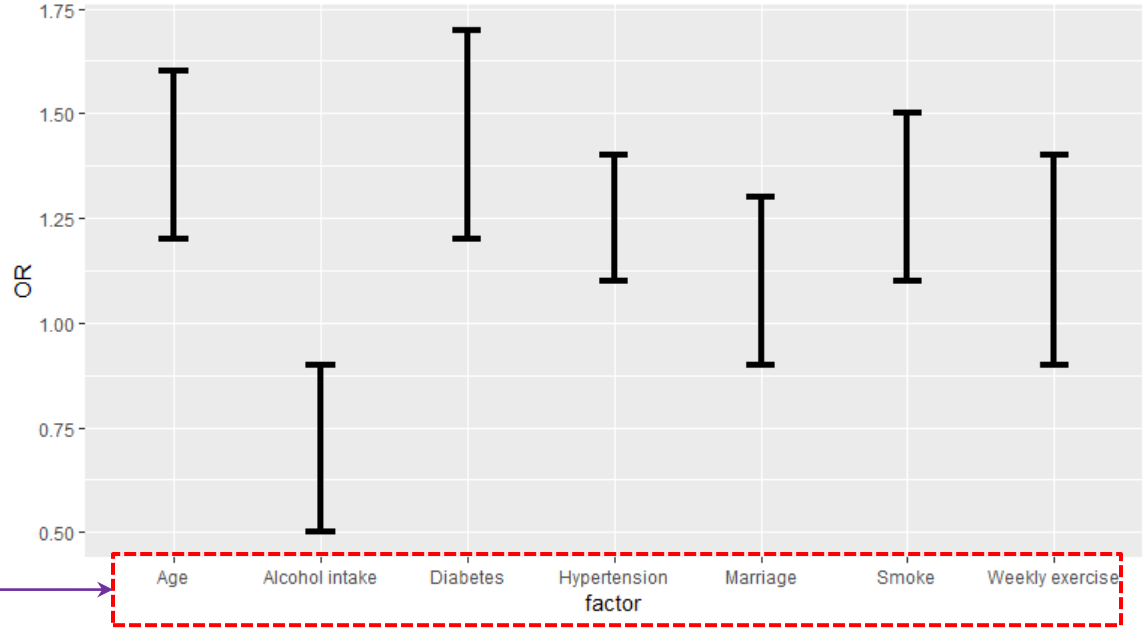
coord_flip()



Thực hành

forestPlot %>%

```
ggplot(aes(x = factor, y = OR, ymin = low, ymax = high)) +
geom_errorbar(width=.2, size = 1.5, show.legend = F)
```



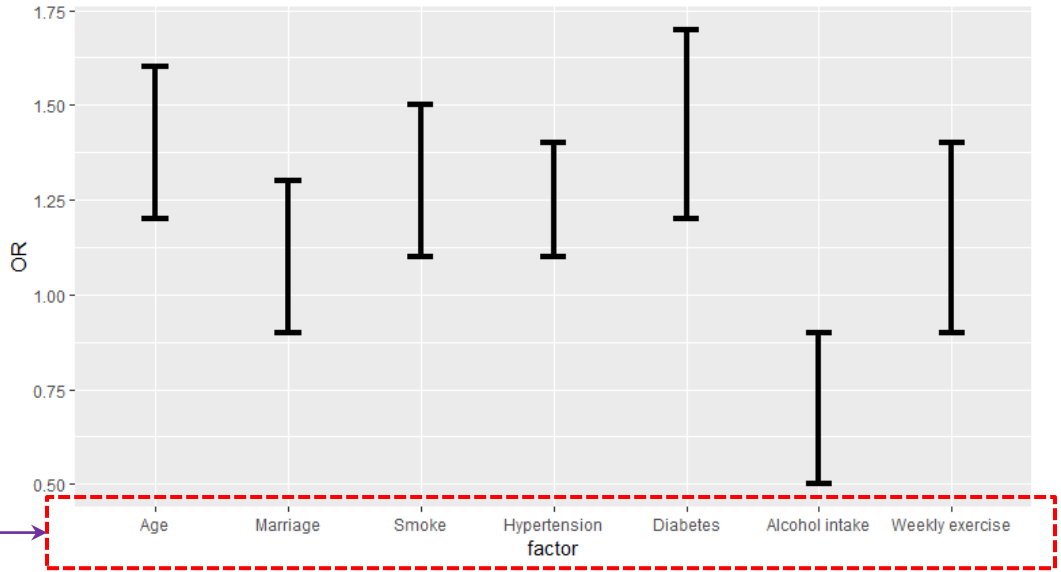
Mặc định xếp theo alphabet

Xắp xếp lộn xộn

Tricks!

Chuyển biến “factor” thành dạng factor và thay đổi level của nó theo thứ tự mong muốn

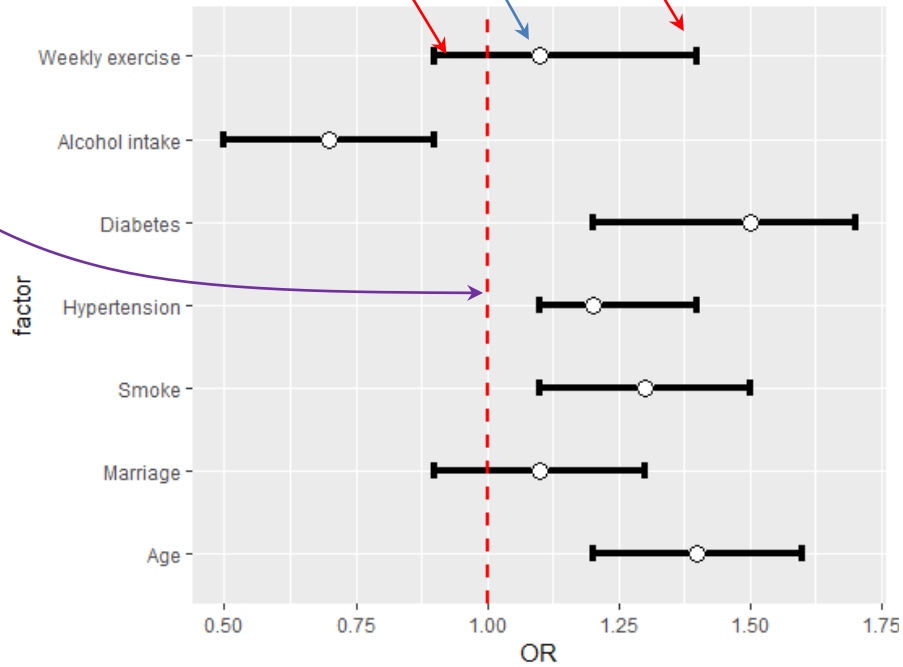
```
forestPlot$factor <- factor(forestPlot$factor,
                            levels = c("Age","Marriage", "Smoke", "Hypertension",
                                       "Diabetes", "Alcohol intake", "Weekly exercise"))
```



Xắp xếp theo ý

Thực hành

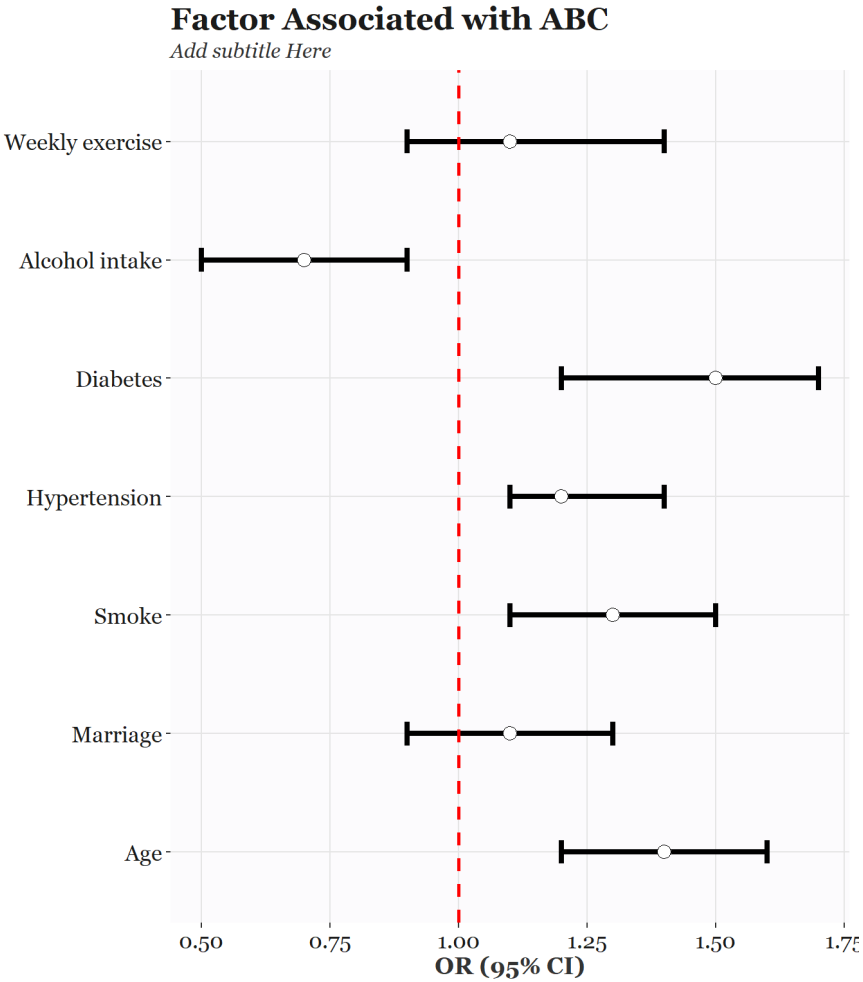
```
forestPlot %>%
        ggplot(aes(x = factor, y = OR, ymin = low, ymax = high)) +
        geom_errorbar(width=.2, size = 1.5, show.legend = F) +
        geom_point(size= 3.5, shape=21, fill="white", show.legend = F)+
        geom_hline(yintercept = 1, linetype = 2, col = "red", size = 1) +
        coord_flip()
```



```r
h1 <- forestPlot %>%
        ggplot(aes(x = factor, y = OR, ymin = low, ymax = high)) +
        geom_errorbar(width=.2, size = 1.5, show.legend = F)+
        geom_point(size= 3.5, shape=21, fill="white", show.legend = F) +
        geom_hline(yintercept = 1, linetype = 2, col = "red", size = 1) +
        coord_flip() +
        labs(title = "Factor Associated with ABC", subtitle = "Add subtitle Here",
                x = "Factors", y = "OR (95% CI)") +
         theme(
                # Chọn font chữ
                text = element_text(family = "Georgia"),
                # Tùy chỉnh text cho title (cỡ chữ 18, bold)
                plot.title = element_text(size = 18,color = "grey10", face = "bold"),
                # Tùy chỉnh cho subtitle
                plot.subtitle = element_text(face = "italic", color = "gray20", size = 12),
                # Tùy chỉnh cation
                plot.caption = element_text(face = "italic", size = 12, color = "gray40"),
                # Tùy chỉnh title cho trục x
                axis.title.x = element_text(face = "bold", size = 14, color = "grey20"),
                # Tùy chinh title cho trục y
                axis.title.y = element_blank(),
                # Tùy chỉnh background, grid
                panel.grid.major = element_line(color = "gray90"),
                #panel.grid.major.y = element_blank(),
                panel.grid.minor.y = element_blank(),
                panel.grid.minor.x = element_blank(),
                panel.background = element_rect(fill = "#fcfbfd"),
                axis.text.x = element_text(size = 13, color = "gray10"),
                axis.text.y = element_text(size = 13, color = "gray10") )
```

Save

```
ggsave("forest_plot.png", h1,
       height = 8, width = 7, units = "in")
```



Trích xuất Coef từ model

Data

```
df <- read_csv("https://gitlab.com/LongKhuong/bayesian_hanoi/raw/master/birthweight.csv")
df %<>% mutate(nghenghiep = as.factor(nghenghiep), tuoithai_center = tuoithai - 38)
df %>% head()
```

```
## # A tibble: 6 x 9
##    maso tuoime tang_ha tuoithai  gioi tlsosinh nghenghiep nhecan
##   <int>  <int>   <int>    <dbl> <int>    <int> <fct>       <int>
## 1     1     33       0     37.7     0     2410 3               1
## 2     2     34       0     39.2     0     2977 2               0
## 3     3     34       0     35.7     0     2100 2               1
## 4     4     30       0     39.3     1     3270 3               0
## 5     5     35       0     38.4     0     2620 2               0
## 6     6     37       0     37.9     1     3260 3               0
## # ... with 1 more variable: tuoithai_center <dbl>
```

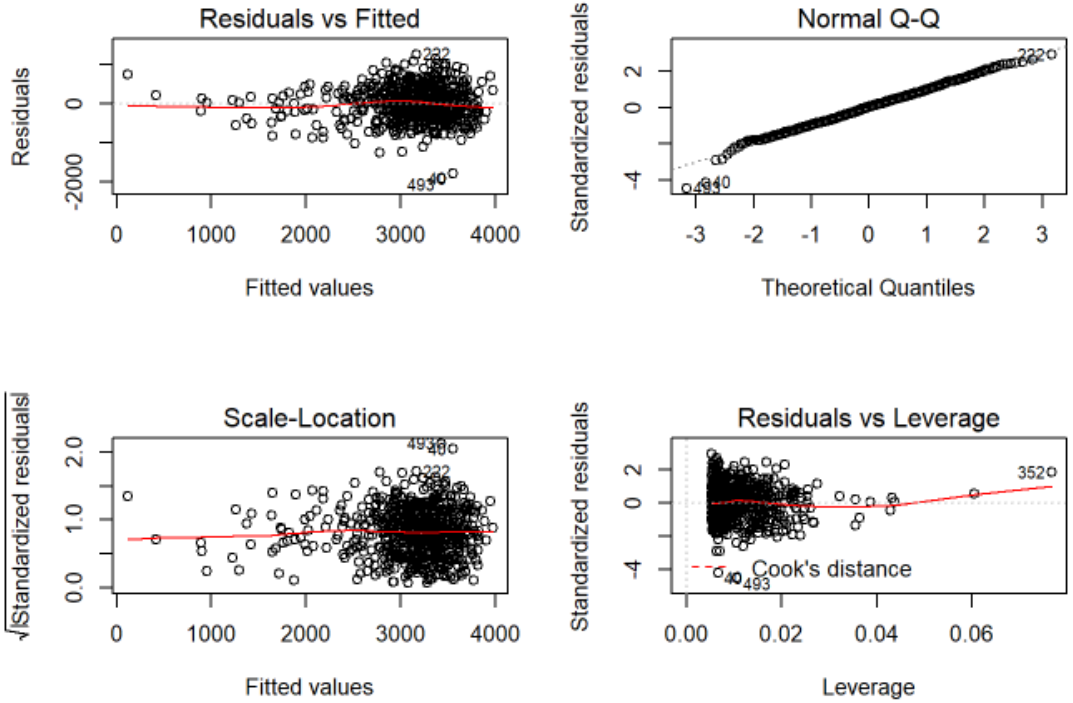
Build model

```
linear_model <- lm(tlsosinh ~ tuoime + tang_ha + tuoithai_center + gioi + nghenghiep, data = df)
summary(linear_model)
```

```
##
## Call:
## lm(formula = tlsosinh ~ tuoime + tang_ha + tuoithai_center +
##     gioi + nghenghiep, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1907.53  -295.95     7.84   277.17  1245.85
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2723.385    156.091  17.447  < 2e-16 ***
## tuoime             1.715      4.396   0.390 0.696521
## tang_ha         -141.483     50.674  -2.792 0.005396 **
## tuoithai_center  201.108      7.486  26.865  < 2e-16 ***
## gioi             166.103     33.946   4.893 1.26e-06 ***
## nghenghiep2      156.183     50.315   3.104 0.001994 **
## nghenghiep3      185.334     48.716   3.804 0.000156 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 427.7 on 634 degrees of freedom
## Multiple R-squared:  0.5747, Adjusted R-squared:  0.5706
## F-statistic: 142.8 on 6 and 634 DF,  p-value: < 2.2e-16
```

Diagnostic

```
par(mfrow=c(2,2))
plot(linear_model)
```



Trích xuất thông tin model

Đầu vào của ggplot2 phải là dataframe ➔ cần chuyển kết quả mô hình thành dataframe

Sử dụng package “broom”

1. tidy : trích xuất nội dung mô hình (Coef, SE, P-value..)
2. glance: trích xuất model fit (AIC, BIC, Rsq…)
3. augment: cung cấp fitted index, residuals…
4. bootstrap: hỗ trợ bootstrap cho gần như tất cả các mô hình

Trích xuất thông tin model

```
library(broom)
linear_summary <- tidy(linear_model)
linear_summary
```

```
## # A tibble: 7 x 5
##   term            estimate std.error statistic   p.value
##   <chr>              <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      2723.      156.      17.4   5.72e- 56
## 2 tuoime              1.72      4.40     0.390 6.97e-  1
## 3 tang_ha          -141.       50.7     -2.79  5.40e-  3
## 4 tuoithai_center   201.        7.49    26.9   9.97e-107
## 5 gioi              166.       33.9      4.89  1.26e-  6
## 6 nghenghiep2       156.       50.3      3.10  1.99e-  3
## 7 nghenghiep3       185.       48.7      3.80  1.56e-  4
```

Trích xuất thông tin model

```
# tính ktc 95%
linear_summary %<>% mutate(high = estimate + 1.96*std.error, low = estimate - 1.96*std.error)
# Bỏ dòng thứ nhất (intercept)
linear_summary %<>% select(term, estimate, low, high) %>% slice(-1)
linear_summary %>% head()
```

```
## # A tibble: 6 x 4
##   term            estimate     low   high
##   <chr>              <dbl>   <dbl>  <dbl>
## 1 tuoime              1.72   -6.90   10.3
## 2 tang_ha          -141.    -241.    -42.2
## 3 tuoithai_center   201.     186.    216.
## 4 gioi              166.      99.6   233.
## 5 nghenghiep2       156.      57.6   255.
## 6 nghenghiep3       185.      89.9   281.
```

Cần điều chỉnh label value của “term” để hiển thị trên biểu đồ

Trích xuất thông tin model

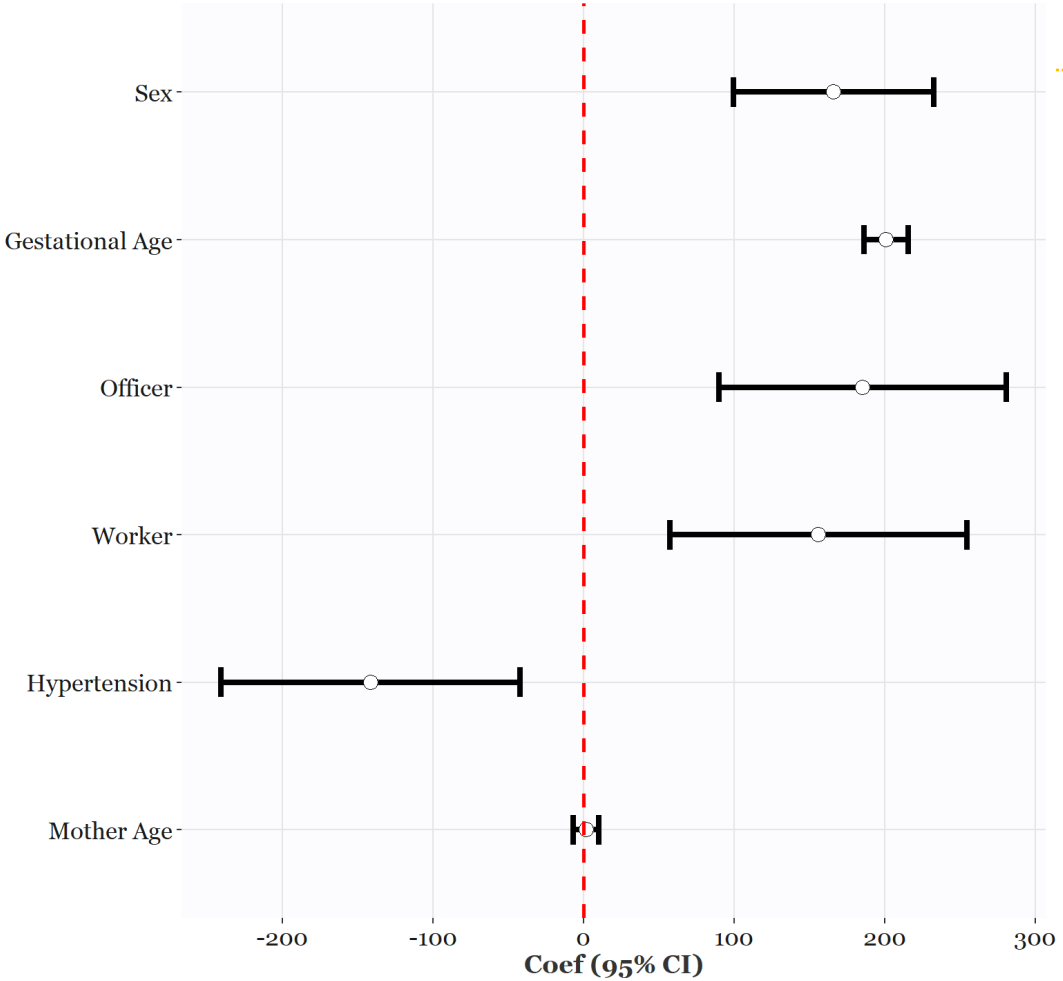
```
# Đổi nhãn giá trị (vẫn giữ nguyên dạng chr)
linear_summary %<>% mutate(term = case_when(term == "tuoime" ~ "Mother Age",
                                            term == "tang_ha" ~ "Hypertension",
                                            term == "nghenghiep2" ~ "Worker",
                                            term == "nghenghiep3" ~ "Officer",
                                            term == "tuoithai_center" ~ "Gestational Age",
                                            term == "gioi" ~ "Sex"))

# đổi level ➔ sắp xếp trên biểu đồ theo ý
linear_summary$term <- factor(linear_summary$term, levels = c("Mother Age", "Hypertension",
"Worker", "Officer", "Gestational Age", "Sex"))
linear_summary
```

```
## # A tibble: 6 x 4
##   term            estimate     low   high
##   <fct>              <dbl>   <dbl>  <dbl>
## 1 Mother Age          1.72   -6.90   10.3
## 2 Hypertension     -141.    -241.    -42.2
## 3 Gestational Age   201.     186.    216.
## 4 Sex               166.      99.6   233.
## 5 Worker            156.      57.6   255.
## 6 Officer           185.      89.9   281.
```

```r
linear_summary %>%
        ggplot(aes(x = term, y = estimate, ymin = low, ymax = high)) +
        geom_errorbar(width=.2, size = 1.5, show.legend = F)+
        geom_point(size= 3.5, shape=21, fill="white", show.legend = F)+
        geom_hline(yintercept = 0, linetype = 2, col = "red", size = 1)+
        coord_flip() +
        labs(title = "Factor Associated with Birth Weight", subtitle = "Add subtitle Here", x = "Factors", y = "Coef (95% CI)") +
        theme(
         # Chọn font chữ
        text = element_text(family = "Georgia"),
        # Tùy chỉnh text cho title (cỡ chũ 18, bold)
        plot.title = element_text(size = 18,color = "grey10", face = "bold"),
        # Tùy chỉnh cho subtitle
        plot.subtitle = element_text(face = "italic", color = "gray20", size = 12),
        # Tùy chỉnh cation
        plot.caption = element_text(face = "italic", size = 12, color = "gray40"),
        # Tùy chỉnh title cho trục x
        axis.title.x = element_text(face = "bold", size = 14, color = "grey20"),
        # Tùy chinh title cho trục y
        axis.title.y = element_blank(),
        # Tùy chỉnh background, grid
        panel.grid.major = element_line(color = "gray90"), #panel.grid.major.y = element_blank(),
        panel.grid.minor.y = element_blank(),
        panel.grid.minor.x = element_blank(),
        panel.background = element_rect(fill = "#fcfbfd"),
        axis.text.x = element_text(size = 13, color = "gray10"),
        axis.text.y = element_text(size = 13, color = "gray10") )
```

Factor Associated with Birth Weight

Add subtitle Here



Ví dụ với logistic model

```
logit_model <- glm(nhecan ~ tuoime + tang_ha + tuoithai_center + gioi + nghenghiep, data = df, family = binomial(link = "logit"))
logit_summry <- tidy(logit_model)
```

```
logit_summry %<>% mutate(OR = exp(estimate),
                         high = exp(estimate + 1.96*std.error),
                         low = exp(estimate - 1.96*std.error)) %>%
        select(term, OR, low, high) %>% slice(-1)
```

```
logit_summry %<>% mutate(term = case_when(term == "tuoime" ~ "Mother Age",
                                          term == "tang_ha" ~ "Hypertension",
                                          term == "nghenghiep2" ~ "Worker",
                                          term == "nghenghiep3" ~ "Officer",
                                          term == "tuoithai_center" ~ "Gestational Age",
                                          term == "gioi" ~ "Sex"))
logit_summry$term <- factor(logit_summry$term, levels = c("Mother Age", "Hypertension", "Worker", "Officer", "Gestational Age", "Sex"))
logit_summry
```

```
## # A tibble: 6 x 4
##   term               OR   low  high
##   <fct>           <dbl> <dbl> <dbl>
## 1 Mother Age      0.971 0.890 1.06
## 2 Hypertension    2.56  1.17  5.59
## 3 Gestational Age 0.388 0.320 0.472
## 4 Sex             0.631 0.319 1.25
## 5 Worker          0.421 0.165 1.08
## 6 Officer         0.313 0.123 0.801
```

MAP

Package “raster”

- raster là một package rất mạnh trong spatial analysis, cũng rất hữu ích trong việc thu thập số liệu
- Trong bài này tập trung vào khả năng thu thập số liệu với **getData()** function

getData()

1. Global admin boundaries (GADM)
2. World Climate data (worldclim)
3. SRTM 90 Data (SRTM)

getData()

```
library(raster)
vietnam_province <- getData("GADM", country = "VNM", level = 1)
# Kiểm tra dạng dữ liệu lấy về từ getData()
vietnam_province %>% class()
```

```
## [1] "SpatialPolygonsDataFrame"
## attr(,"package")
## [1] "sp"
```

Level 0: Quốc gia
Level 1: Tỉnh
Level 2: Huyện

Mã A3***

Cần chuyển dữ liệu về dạng data frame dựa vào function "fortify()" trong gói **ggplot2**

```
vietnam_df <- vietnam_province %>% fortify(region = "NAME_1") %>% as.tibble()
```

*** tham khảo mã A3 tại: http://kirste.userpage.fu-berlin.de/diverse/doc/ISO_3166.html

- Tuy nhiên, trước khi sử dụng được function fortify này, cần chuyển status của gói **gpclib** sang **“TRUE”**
- Note: Nếu status chưa **“TRUE”** sẽ thấy những dòng error này

```
    Error in maptools::unionSpatialPolygons(cp,
attr[, region]) :
    isTRUE(gpclibPermitStatus()) is not TRUE
```

Khi đó cần chạy các lệnh này

```
library(rgdal)
library(maptools)
if (!require(gpclib)) install.packages("gpclib", type="source")
gpclibPermit()
```

```
## [1] TRUE
```

```
vietnam_df <- vietnam_province %>% fortify(region = "NAME_1") %>% as.tibble()
vietnam_df %>% head()
```

```
## # A tibble: 6 x 7
##    long   lat order hole  piece id       group
##   <dbl> <dbl> <int> <lgl> <fct> <chr>    <fct>
## 1  105.  10.2     1 FALSE 1     An Giang An Giang.1
## 2  105.  10.2     2 FALSE 1     An Giang An Giang.1
## 3  105.  10.3     3 FALSE 1     An Giang An Giang.1
## 4  105.  10.3     4 FALSE 1     An Giang An Giang.1
## 5  105.  10.3     5 FALSE 1     An Giang An Giang.1
## 6  105.  10.3     6 FALSE 1     An Giang An Giang.1
```

```
# Xem tên các tỉnh
vietnam_df$id %>% unique()
```

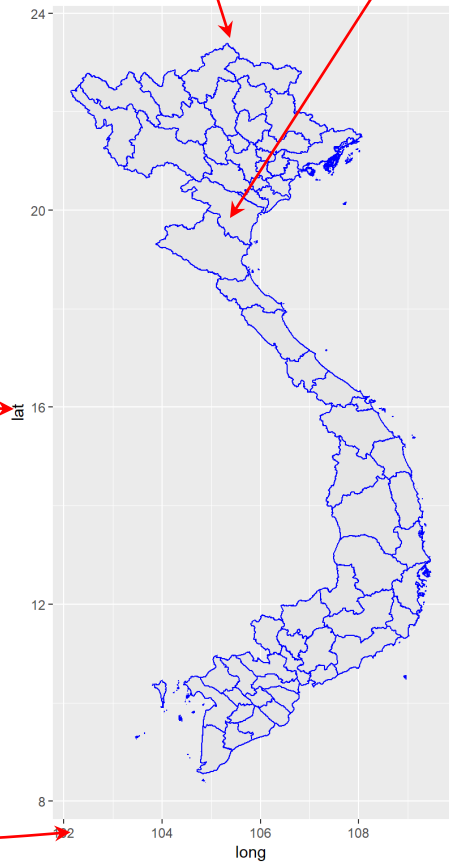
```
##  [1] "An Giang"          "B<U+1EA1>c Liêu"   "B<U+1EAF>c Giang"
##  [4] "B<U+1EAF>c K<U+1EA1>n" "B<U+1EAF>c Ninh"   "B<U+1EBF>n Tre"
##  [7] "Bà R<U+1ECB>a - Vung Tàu" "Bình Đ<U+1ECB>nh"  "Bình Duong"
## [10] "Bình Phu<U+1EDB>c" "Bình Thu<U+1EAD>n" "C<U+1EA7>n Tho"
## [13] "Cà Mau"            "Cao B<U+1EB1>ng"   "Đ<U+1EAF>k L<U+1EAF>k"
## [16] "Đ<U+1EAF>k Nông"   "Đ<U+1ED3>ng Nai"   "Đ<U+1ED3>ng Tháp"
## [19] "Đà N<U+1EB5>ng"    "Đi<U+1EC7>n Biên"  "Gia Lai"
## [22] "H<U+1EA3>i Duong"  "H<U+1EA3>i Phòng"  "H<U+1EAD>u Giang"
## [25] "H<U+1ED3> Chí Minh" "Hà Giang"          "Hà N<U+1ED9>i"
## [28] "Hà Nam"            "Hà Tinh"           "Hoà Bình"
## [31] "Hung Yên"          "Khánh Hòa"         "Kiên Giang"
## [34] "Kon Tum"           "L<U+1EA1>ng Son"   "Lai Châu"
## [37] "Lâm Đ<U+1ED3>ng"   "Lào Cai"           "Long An"
## [40] "Nam Đ<U+1ECB>nh"   "Ngh<U+1EC7> An"    "Ninh Bình"
## [43] "Ninh Thu<U+1EAD>n" "Phú Th<U+1ECD>"    "Phú Yên"
## [46] "Qu<U+1EA3>ng Bình" "Qu<U+1EA3>ng Nam"  "Qu<U+1EA3>ng Ngãi"
## [49] "Qu<U+1EA3>ng Ninh" "Qu<U+1EA3>ng Tr<U+1ECB>" "Sóc Trang"
## [52] "Son La"            "Tây Ninh"          "Th<U+1EEB>a Thiên Hu<U+1EBF>"
## [55] "Thái Bình"         "Thái Nguyên"       "Thanh Hóa"
## [58] "Ti<U+1EC1>n Giang" "Trà Vinh"          "Tuyên Quang"
## [61] "Vinh Long"         "Vinh Phúc"         "Yên Bái"
```

- Các tỉnh đang sử dụng chữ **Tiếng Việt có dấu**, rất dễ lỗi font và khó merge với data khác sau này ==> chuyển hết font về định dạng **Latin-ASCII**.
- Có thể chuyển định dạng này bằng function "stri_trans_general()" trong package **stringi**

```
# install.packages("stringi")
library(stringi)
vietnam_df %<>% mutate(id = stri_trans_general(id, "Latin-ASCII"), group =
stri_trans_general(group,"Latin-ASCII"))
# xem lai ten
vietnam_df$id %>% unique()
# Lưu lại "Shape file" có thể sử dụng về sau, không phải tải lại
write.csv(vietnam_df , "shape_file_vn.csv", na = "")
```

```
##  [1] "An Giang"          "Bac Lieu"          "Bac Giang"
##  [4] "Bac Kan"           "Bac Ninh"          "Ben Tre"
##  [7] "Ba Ria - Vung Tau" "Binh Dinh"         "Binh Duong"
## [10] "Binh Phuoc"        "Binh Thuan"        "Can Tho"
## [13] "Ca Mau"            "Cao Bang"          "Dak Lak"
## [16] "Dak Nong"          "Dong Nai"          "Dong Thap"
## [19] "Da Nang"           "Dien Bien"         "Gia Lai"
## [22] "Hai Duong"         "Hai Phong"         "Hau Giang"
## [25] "Ho Chi Minh"       "Ha Giang"          "Ha Noi"
## [28] "Ha Nam"            "Ha Tinh"           "Hoa Binh"
## [31] "Hung Yen"          "Khanh Hoa"         "Kien Giang"
## [34] "Kon Tum"           "Lang Son"          "Lai Chau"
## [37] "Lam Dong"          "Lao Cai"           "Long An"
## [40] "Nam Dinh"          "Nghe An"           "Ninh Binh"
## [43] "Ninh Thuan"        "Phu Tho"           "Phu Yen"
## [46] "Quang Binh"        "Quang Nam"         "Quang Ngai"
## [49] "Quang Ninh"        "Quang Tri"         "Soc Trang"
## [52] "Son La"            "Tay Ninh"          "Thua Thien Hue"
## [55] "Thai Binh"         "Thai Nguyen"       "Thanh Hoa"
## [58] "Tien Giang"        "Tra Vinh"          "Tuyen Quang"
## [61] "Vinh Long"         "Vinh Phuc"         "Yen Bai"
```

```
vietnam_df %>% ggplot() +
geom_polygon(aes(x = long, y = lat, group = group), color = "blue", fill = "grey90") +
coord_equal()
```



fill sẽ sử dụng để "fill" số liệu về sau

Tóm tắt

```
library(raster)
vietnam_province <- getData("GADM", country = "VNM", level = 1)
# Chạy lệnh để dùng được function fortify()
library(rgdal)
library(maptools)
if (!require(gpclib)) install.packages("gpclib", type="source")
gpclibPermit()
# chuyển về data frame bằng function fortify()
vietnam_df <- vietnam_province %>% fortify(region = "NAME_1") %>% as.tibble()
# Chuyển font tên tỉnh để merge số liệu về sau
library(stringi)
vietnam_df %<>% mutate(id = stri_trans_general(id, "Latin-ASCII"), group =
stri_trans_general(group,"Latin-ASCII"))
# Vẽ thôi!
vietnam_df %>% ggplot() +
geom_polygon(aes(x = long, y = lat, group = group), color = "blue", fill = "grey90") +
coord_equal()
```

Merge với số liệu cụ thể

Phương pháp

- Có số liệu tương ứng với tên của (các tỉnh nếu map level 1, các huyện nếu map level 2).
- Điều chỉnh tên (tên tỉnh/huyện) giữa shape file và dữ liệu cần fill
- Merge 2 file theo tên
- Vẽ map và fill dữ liệu

Data

```
library(readxl)
# dữ liệu trong file excel, giá trị missing được đánh dấu bằng dấu - . ➔ phải khai báo na = "-“
province <- read_xlsx("province.xlsx", na = "-")
head(province)
```

```
## # A tibble: 6 x 14
##    Year City    Total `0-14` `15-64` `65+` `Female 15-49` `Median age`
##   <dbl> <chr>   <chr> <chr>  <chr>   <chr> <chr>                  <dbl>
## 1  2009 Ha Noi 6 452 1 428  4 566   457   1 909                   28.5
## 2  2014 Ha Noi 7 019 1 645  4 881   493   1 946                   29.9
## 3  2019 Ha Noi 7 494 1 835  5 084   575   1 990                   32
## 4  2024 Ha Noi 7 875 1 834  5 290   750   2 053                   34.2
## 5  2029 Ha Noi 8 155 1 699  5 499   957   2 121                   36.4
## 6  2034 Ha Noi 8 383 1 569  5 745   1 069 2 121                   37.8
## # ... with 6 more variables: `Age dependency ratio` <dbl>, `Aging
## #   index` <dbl>, `Sex ratio` <dbl>, `Crude birth rate` <dbl>, `Crude
## #   death rate` <dbl>, `Natural growth rate` <dbl>
```

Mục tiêu: vẽ tỷ lệ nhóm tuổi 65+ theo các tỉnh và facet theo 4 năm 2019, 2024, 2029 và 2034

Cleaning data

head(province)


```
## # A tibble: 6 x 14
##    Year City   Total `0-14` `15-64` `65+` `Female 15-49` `Median age`
##   <dbl> <chr>  <chr> <chr>  <chr>   <chr> <chr>                  <dbl>
## 1  2009 Ha Noi 6 452 1 428  4 566   457   1 909                   28.5
## 2  2014 Ha Noi 7 019 1 645  4 881   493   1 946                   29.9
## 3  2019 Ha Noi 7 494 1 835  5 084   575   1 990                   32
## 4  2024 Ha Noi 7 875 1 834  5 290   750   2 053                   34.2
## 5  2029 Ha Noi 8 155 1 699  5 499   957   2 121                   36.4
## 6  2034 Ha Noi 8 383 1 569  5 745   1 069 2 121                   37.8
## # ... with 6 more variables: `Age dependency ratio` <dbl>, `Aging
## #   index` <dbl>, `Sex ratio` <dbl>, `Crude birth rate` <dbl>, `Crude
## #   death rate` <dbl>, `Natural growth rate` <dbl>
```

```r
# Đổi tên các biến số cho thống nhất
names(province) <- c("year", "id", "total", "age0_14", "age15_64", "age65", "female15_49",
"age_median", "age_dependency_ratio", "aging_index", "sex_ratio", "crude_birth_rate",
"crude_death_rate", "natural_growth_rate")
# Xem cấu trúc dữ liệu
glimpse(province)
```

```
## Observations: 378
## Variables: 14
## $ year                <dbl> 2009, 2014, 2019, 2024, 2029, 2034, 2009,...
## $ id                  <chr> "Ha Noi", "Ha Noi", "Ha Noi", "Ha Noi", "...
## $ total               <chr> "6 452", "7 019", "7 494", "7 875", "8 15...
## $ age0_14             <chr> "1 428", "1 645", "1 835", "1 834", "1 69...
## $ age15_64            <chr> "4 566", "4 881", "5 084", "5 290", "5 49...
## $ age65               <chr> "457", "493", "575", "750", "957", "1 069...
## $ female15_49         <chr> "1 909", "1 946", "1 990", "2 053", "2 12...
## $ age_median          <dbl> 28.5, 29.9, 32.0, 34.2, 36.4, 37.8, 23.1,...
## $ age_dependency_ratio <dbl> 41.3, 43.8, 47.4, 48.9, 48.3, 45.9, 58.3,...
## $ aging_index         <dbl> 44.4, 44.1, 50.2, 63.1, 76.8, 94.0, 19.6,...
## $ sex_ratio           <dbl> 96.6, 97.3, 97.6, 97.7, 97.7, 97.5, 100.3...
## $ crude_birth_rate    <dbl> NA, 19.1, 17.6, 14.9, 12.8, 12.2, NA, 22....
## $ crude_death_rate    <dbl> NA, 6.2, 6.2, 6.2, 6.6, 7.3, NA, 7.2, 6.6...
## $ natural_growth_rate <dbl> NA, 12.9, 11.4, 8.7, 6.1, 4.9, NA, 15.4, ...
```

Dạng <chr>??

Vì có dấu cách giữa đơn vị hàng trăm và ngàn ➔ loại bỏ dấu cách và chuyển về numeric

Destring

Có 2 cách

1. Sử dụng 1 for loop

```
# Vì biến total :female15-49 có index từ 3:7
for (i in 3:7) {
        # thay đổi " " bằng "", sau đó chuyển về numeric
        province[[i]] <- province[[i]] %>% str_replace(" ", "") %>% as.numeric()
}
head(province)
```

2. Sử dụng 1 function + mutate()

```
destring <- function(x) {
        x %>% str_replace(" ", "") %>%  as.numeric()
}
province %<>% mutate(total = destring(total),
        age0_14 = destring(age0_14),
         age15_64 = destring(age15_64),
        age65 = destring(age65),
        female15_49 = destring(female15_49))
```

```
head(province)
```

```
## # A tibble: 6 x 14
##    year id     total age0_14 age15_64 age65 female15_49 age_median
##   <dbl> <chr>  <dbl>   <dbl>    <dbl> <dbl>       <dbl>      <dbl>
## 1  2009 Ha Noi  6452    1428     4566   457        1909       28.5
## 2  2014 Ha Noi  7019    1645     4881   493        1946       29.9
## 3  2019 Ha Noi  7494    1835     5084   575        1990       32
## 4  2024 Ha Noi  7875    1834     5290   750        2053       34.2
## 5  2029 Ha Noi  8155    1699     5499   957        2121       36.4
## 6  2034 Ha Noi  8383    1569     5745  1069        2121       37.8
## # ... with 6 more variables: age_dependency_ratio <dbl>,
## #   aging_index <dbl>, sex_ratio <dbl>, crude_birth_rate <dbl>,
## #   crude_death_rate <dbl>, natural_growth_rate <dbl>
```

Các biên đã đúng định dạng ➔ tạo biên tỷ lệ 65+

```
province %<>% mutate(age65_rate = age65/(age0_14 + age15_64 + age65),
                     age65_rate_group = cut(age65_rate, breaks = c(0, 0.08, 0.1, 1),
                     levels = c(1,2,3), labels = c("<8%", "8-10%", ">10%")))
table(province$age65_rate_group)
```

```
##
##   <8% 8-10%  >10%
##   194    75   109
```

Chuyển id (tên tỉnh) về định dạng "Latin-ASCII" sau đó điều chỉnh để trùng với shape file

```
province %<>% mutate(id = stri_trans_general(id, "Latin-ASCII"))
# So sánh tên giữa data và shape file
setdiff(province$id, vietnam_df$id)
# có 2 tỉnh ở trong data provine khác với shape file ➔ cần điều chỉnh
```

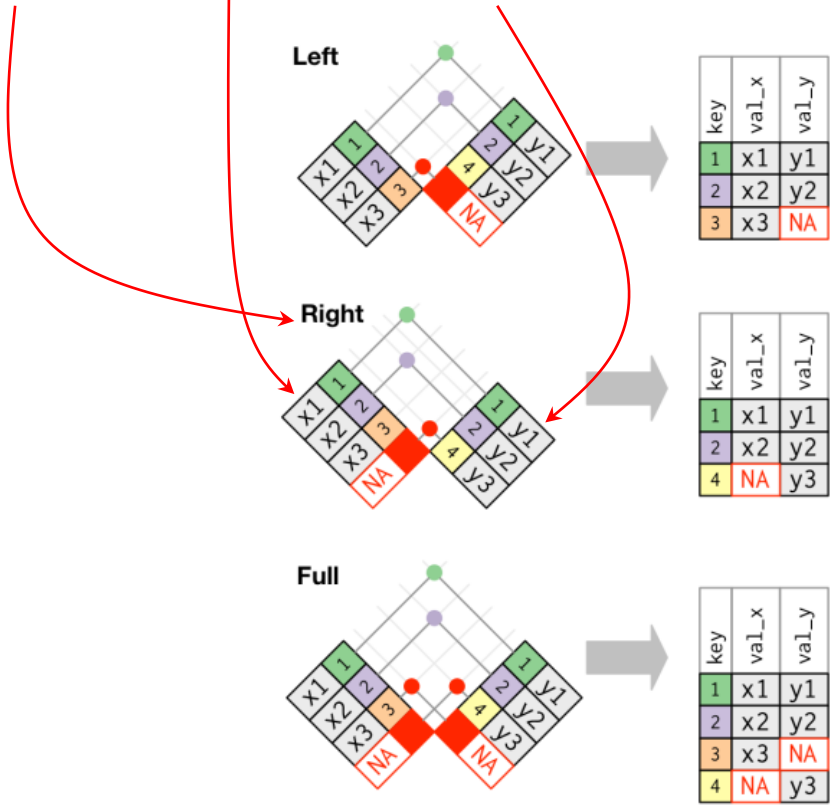
```
## [1] "Ba Ria Vung Tau"  "Ho Chi Minh city"
```

```
# Điều chỉnh id của provine theo đúng giá trị trong shape file
province %<>% mutate(id = case_when(
                                    id == "Ba Ria Vung Tau" ~ "Ba Ria - Vung Tau",
                                    id == "Ho Chi Minh city" ~ "Ho Chi Minh",
                                    TRUE ~ id))

# Check lại lần nữa
setdiff(province$id, vietnam_df$id)
# đã phù hợp 100%
```

```
## character(0)
```

Merge 2 files

```
vietnam_age <- right_join(vietnam_df, province, by = "id")
```



Vẽ map

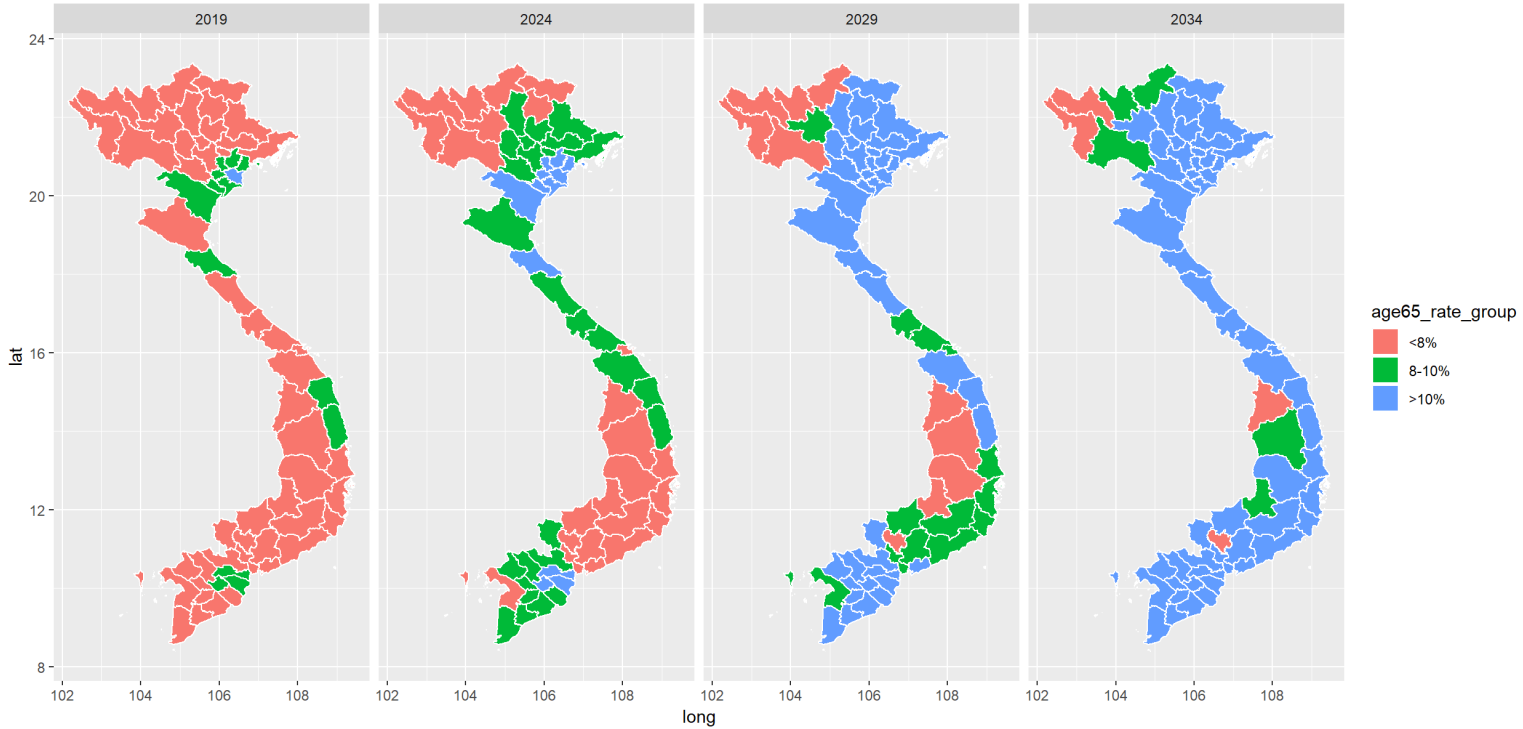
```
table(vietnam_age$year)
```

```
##
##  2009  2014  2019  2024  2029  2034
## 86152 86152 86152 86152 86152 86152
```

Yêu cầu chỉ lấy 4 năm 2019, 2024, 2029, 2034. Do đó phải filter 4 năm này ra bằng cách sử dụng **%in%**

```
# vietnam_age %>% filter(year %in% c(2019, 2024, 2029, 2034))
```

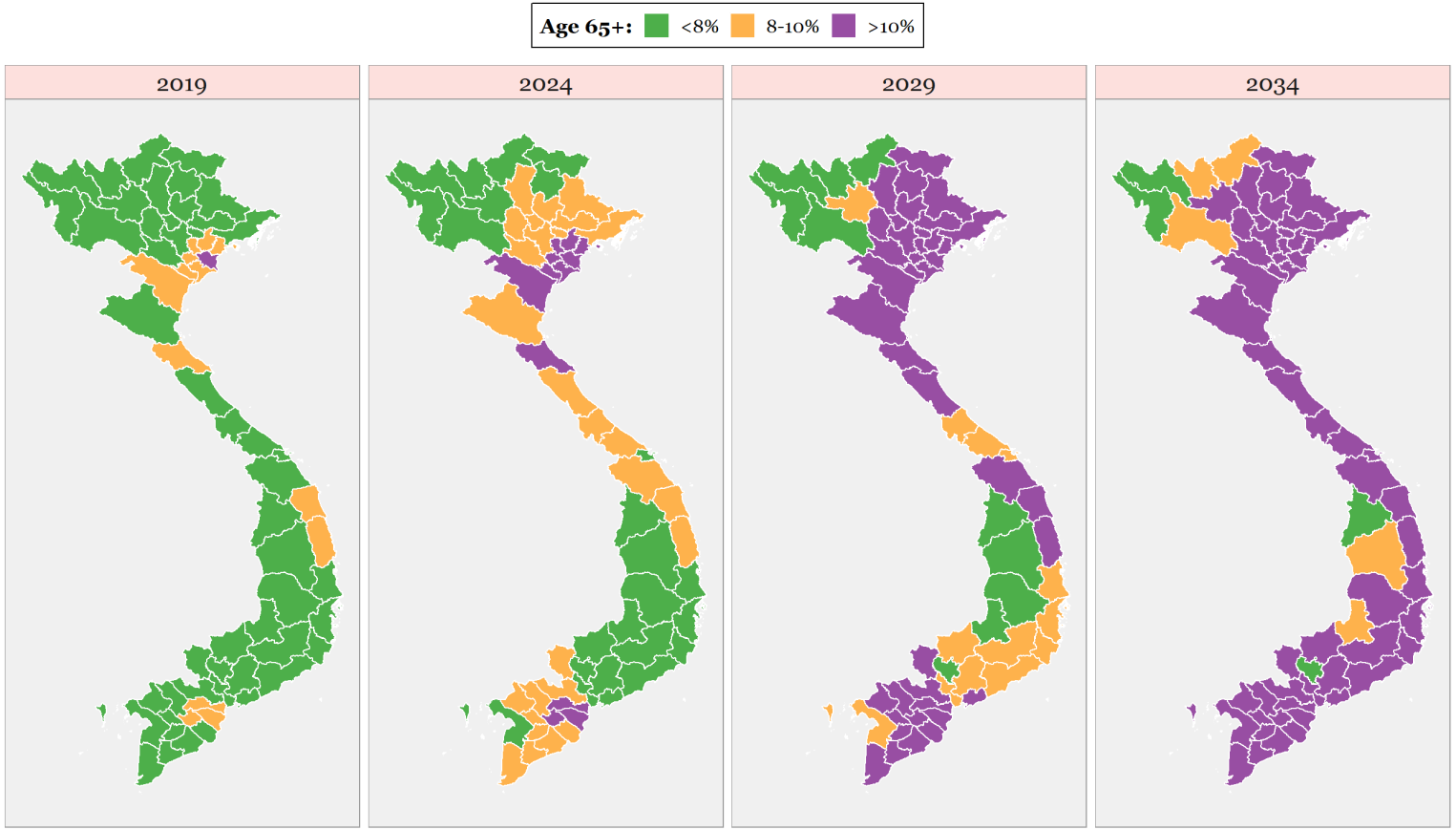
```
vietnam_age %>% filter(year %in% c(2019, 2024, 2029, 2034)) %>%
        ggplot() +
        geom_polygon(aes(x = long, y = lat,  group = group,
                                              fill = age65_rate_group), color = "white", size = 0.5) +
        coord_equal() +
        facet_grid(~ year)
```


2019
2024
2029
2034
24
20
16
12
8
lat
102
104
106
108
long
age65_rate_group
<8%
8-10%
>10%

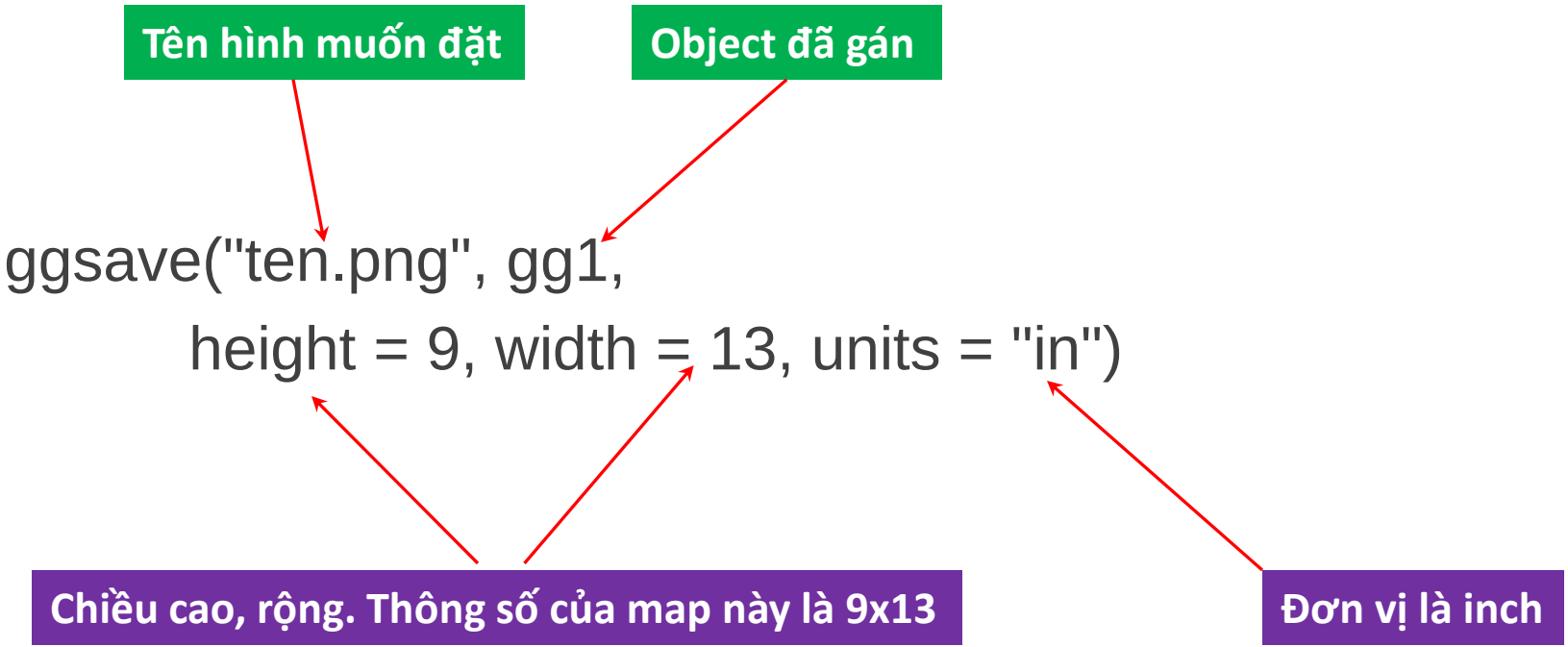
```r
gg1 <- vietnam_age %>% filter(year %in% c(2019, 2024, 2029, 2034)) %>%
  ggplot() +
  geom_polygon(aes(x = long, y = lat, group = group, fill = age65_rate_group),color = "white", size = 0.5) +
  coord_equal() + facet_grid(~ year) +
  scale_fill_manual(name = "Age 65+:", labels = c("<8%", "8-10%", ">10%"),
                    values = c("<8%"= "#4daf4a", "8-10%" = "#feb24c", ">10%" = "#984ea3")) +
  labs( title = "Propotion of Age 65+ in Vietnam by Province from 2019 to 2034",
        subtitle = "Note: Data Is Not Available for\nVietnam's Paracel and Spratly Islands",
        caption = "Data Source: https://www.gso.gov.vn") +
  theme(
  text = element_text(family = "Georgia"),
  plot.title = element_text(size = 14,color = "grey10", face = "bold"),
  plot.subtitle = element_text(face = "italic", color = "gray10", size = 12),
  plot.caption = element_text(face = "italic", size = 10, color = "gray10"),
  axis.line = element_blank(),
  axis.text.x = element_blank(),
  axis.text.y = element_blank(),
  axis.ticks = element_blank(),
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.grid.minor = element_blank(),
  panel.grid.major = element_blank(),
  # hình nền phía ngoài "panel.background"
  plot.background = element_rect(fill = "white", color = "black"),
  # Nên phía trong, hình nền bên rìa hình vẽ chính
  panel.background = element_rect(fill = "#f0f0f0", color = NA),
  # Chinh title cho legend
  legend.title = element_text(size = 13, face = "bold"),
  # Cỡ chữ chú thích, i.e., <8%
  legend.text = element_text(size = 12),
  # Chiều hiển thị legend, i.g., ngang = "horizontal"
  legend.direction = "horizontal", legend.position = "top",
  legend.background = element_rect(fill = NA, color = "black"),
  # viền từng tấm ảnh (giữa các hình)
  panel.border = element_rect(color = "gray60", fill = NA),
  strip.text = element_text(size = 15),
  strip.background = element_rect(color = "gray60", fill = "#fde0dd") )
```

Propotion of Age 65+ in Vietnam by Province from 2019 to 2034

Note: Data Is Not Available for Vietnam's Paracel and Spratly Islands



Data Source: https://www.gso.gov.vn


Tên hình muốn đặt
Object đã gán
ggsave("ten.png", gg1,
height = 9, width = 13, units = "in")
Chiều cao, rộng. Thông số của map này là 9x13
Đơn vị là inch

Thêm các điểm theo tọa độ và anotation

Nguyên lý

- Trục x và y của map tương ứng là longitude & latitude
- Thêm layer như nguyên lý của ggplot2
- Để thêm các marker cần có dữ liệu về long & lat

Trường Sa & Hoàng Sa

- Dữ liệu trên raster không có 2 quần đảo này ➔ thêm 1 layer bổ sung trên map
- Nếu có đầy đủ tất cả các tọa độ của các hòn đảo trong 2 quần đảo này ➔ dùng thêm 1 layer geom_polygon()
- Nếu không có đầy đủ ➔ có thể “dùng tạm” layer geom_point() + annotation

Wraping dữ liệu từ Wikipedia

Trường Sa: https://en.wikipedia.org/wiki/List_of_maritime_features_in_the_Spratly_Islands

Hoàng Sa: https://en.wikipedia.org/wiki/Paracel_Islands

- Đã được lưu tại

Trường Sa: https://gitlab.com/LongKhuong/data-visualization/raw/master/Data/truongsa_geo.csv

Hoàng Sa: https://gitlab.com/LongKhuong/data-visualization/raw/master/Data/hoangsa_geo.csv

- Cần thêm 2 thành phần:

1. Dữ liệu tọa độ của các rải đá (reef) (đã lưu ở gitlab)
2. Tọa độ để đặt 2 labels cho 2 quần đảo này

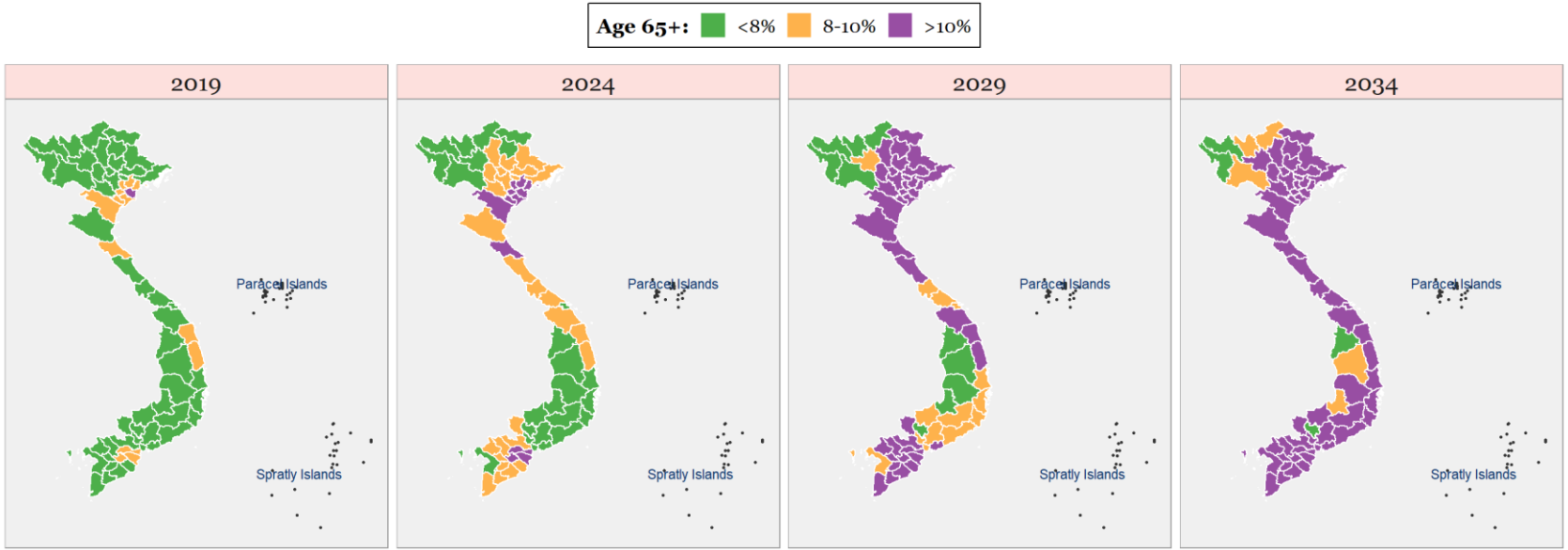
```
ts_hs <- data.frame(lat = c(9.500, 16.96666667),
                    long = c(113.00,112.3333333),
                    name = c("Spratly Islands", "Paracel Islands"))
```

Vẽ

```
# Object đã lưu
gg2 <- gg1 +
            # Hoang Sa
            geom_point(data = hs_geo, aes(x = long, y = lat), size = 0.5, color = "gray20") +
            # Truong Sa
            geom_point(data = ts_geo, aes(x = long, y = lat), size = 0.5, color = "gray20") +
            # Nhãn của 2 quần đảo
            geom_text(data = ts_hs, aes(label = name, x = long, y = lat), size = 3, color =
"#08306b")
gg2
```

Propotion of Age 65+ in Vietnam by Province from 2019 to 2034

Note: Data Is Not Available for Vietnam's Paracel and Spratly Islands



Data Source: https://www.gso.gov.vn

Highlight một số tỉnh

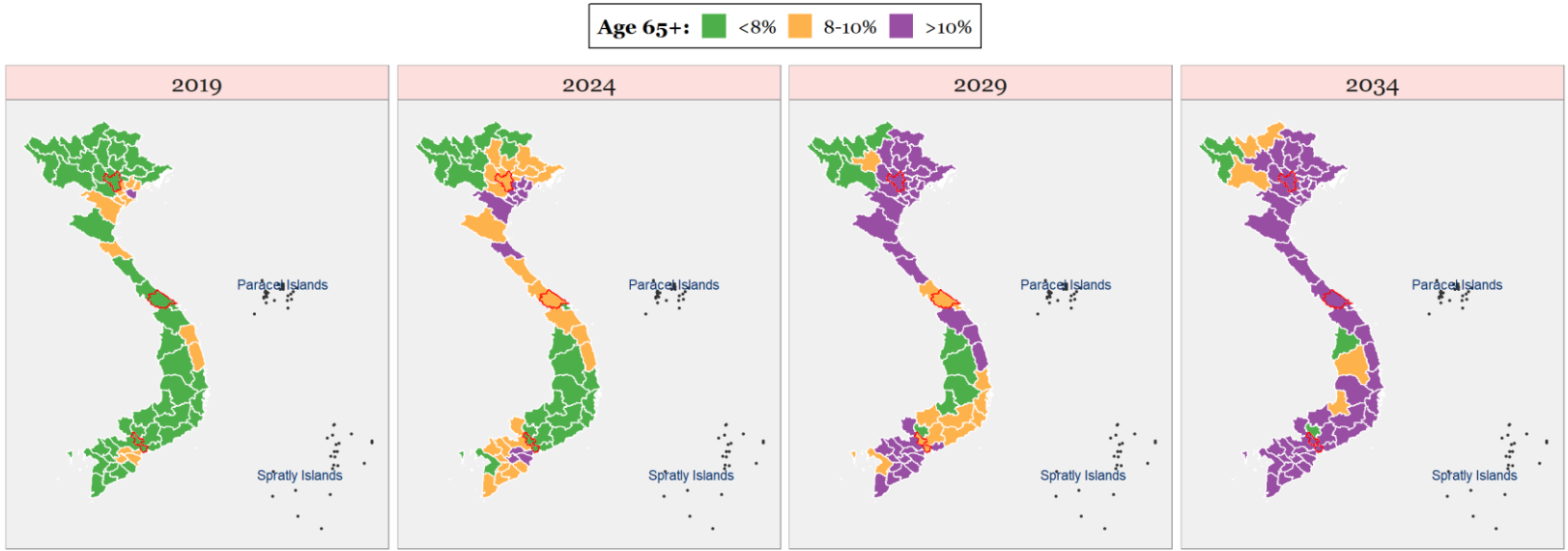
Giả sử cần highlight Hà Nội, Huế và Tp.HCM

- Filter các id cần highlight từ dữ liệu đầy đủ, sử dụng %in%
- Sử dụng geom_path() thay vì geom_polygon() vì chỉ tô viền tỉnh tránh chồng lấp màu

```
gg2 +
        geom_path(data = vietnam_age %>% filter(id %in% c("Ha Noi", "Thua
Thien Hue", "Ho Chi Minh") & year %in% c(2019, 2024, 2029, 2034)) ,
        aes(x = long, y = lat, group = group), color = "red", show.legend = F)
```

Propotion of Age 65+ in Vietnam by Province from 2019 to 2034

Note: Data Is Not Available for Vietnam's Paracel and Spratly Islands



Data Source: https://www.gso.gov.vn

Thank you!