Simple HHI

Acuna

2024-12-06

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(readxl)
```

Import political dynasties data

```r
data <- read_excel("datasets/ASoG-POLITICAL-DYNASTIES-DATASET.XLSX", sheet = "Data")
data[is.na(data)] <- ""
data <- data %>% mutate(Full.Name = paste(FN, LN, sep = " "))
data
```

```
## # A tibble: 86,234 x 10
##    FN         LN    Party Region Province Municipality.City Position  Year   fat
##    <chr>      <chr> <chr> <chr>  <chr>    <chr>             <chr>    <dbl> <dbl>
##  1 BAMER      AAD   KAMPI Auton~ LANAO D~ "CALANOGAS"       COUNCIL~  2007     0
##  2 AMIN       AADAM INDE~ Auton~ LANAO D~ "MAROGONG"        COUNCIL~  2004     0
##  3 AMIN       AADAM LAKA~ Auton~ LANAO D~ "MAROGONG"        COUNCIL~  2007     0
##  4 TRISTAN R~ AALA  IND.  REGIO~ DAVAO D~ "CITY OF TAGUM"   COUNCIL~  2010     1
##  5 SHIRLEY B~ AALA  LAKA~ REGIO~ DAVAO D~ ""                PROVINC~  2004     1
##  6 TRISTAN R~ AALA  LAKA~ REGIO~ DAVAO D~ "CITY OF TAGUM"   COUNCIL~  2004     1
##  7 RAUL       AALA  LAKA~ REGIO~ LAGUNA   "SANTA ROSA CITY" COUNCIL~  2004     0
##  8 MA. THERE~ AALA  LAKA~ REGIO~ LAGUNA   "SANTA ROSA CITY" COUNCIL~  2007     0
##  9 SHIRLEY B~ AALA  LP    REGIO~ DAVAO D~ ""                PROVINC~  2013     1
## 10 TRISTAN R~ AALA  LP    REGIO~ DAVAO D~ ""                COUNCIL~  2013     1
## # i 86,224 more rows
## # i 1 more variable: Full.Name <chr>
```

Create function to compute HHI given the province and year

```r
get_hhi <- function(year, province, region, df){

  # filter by the given region
  filtered <- df %>% filter(Region == region)
  filtered[is.na(filtered)] <- ""

  # filter by the given year and province
  dfX <- filtered %>% filter(Year == year, Province == province)

  if (nrow(dfX) == 0) {
    return(NA)  # Return NA if no data
  }

  # retrieve list of last names
  cols <- unique(dfX$Full.Name)
  col_lns <- sapply(cols, function(col) dfX %>% filter(Full.Name == col) %>% pull(LN) %>% .[1])
  names <- unique(c(col_lns))
  names <- sort(names[names != ""])
  hhi <- data.frame(Family = names)

  # calculate HHI
  scores <- sapply(names, function(ln) sum(dfX$LN == ln))
  hhi$Seats <- scores
  hhi$Seat.Shares <- hhi$Seats / nrow(dfX)

  hh_index <- sum((100 * hhi$Seat.Shares)^2)
  return(hh_index)
}
```

For each year, and then for each province, compute HHI using the `get_hhi` function.

```r
provinces <- unique(data$Province)
regions <- unique(data$Region)
years <- c(2004, 2007, 2010, 2013, 2016)

# initialize empty dataframe
hhi_df <- data.frame(matrix(nrow = length(provinces), ncol = length(years)))
colnames(hhi_df) <- years

for (region in regions) {

  # filter by the given region
  df8 <- data %>% filter(Region == region)
  provinces <- unique(df8$Province)

  # create sub-dataframe for the HHIs in that region
  hhi_a <- data.frame(matrix(nrow = length(provinces), ncol = length(years)))
  colnames(hhi_a) <- years
  rownames(hhi_a) <- provinces

  # compute HHI for each province and year
  for (province in provinces) {
    for (year in years) {
```

```
      hhi_a[province, as.character(year)] <- get_hhi(year, province, region, data)
    }
  }

  # row-bind the sub-dataframe to the main dataframe
  hhi_df <- rbind(hhi_df, hhi_a)
}

hhi_df <- hhi_df[complete.cases(hhi_df),]
hhi_df
```

```
##                                    2004      2007      2010      2013
## LANAO DEL SUR                  46.84711  49.89349  68.73035  74.60973
## MAGUINDANAO                   126.93136 150.12959 152.62346 115.27281
## SULU                           85.17795  91.14583  87.68201 141.89609
## BASILAN                       137.90100 144.79500  90.66358  97.50297
## TAWI-TAWI                     109.56903 109.09091 133.92857 131.82160
## DAVAO DEL NORTE                88.68584  88.38384  90.67952  90.67952
## DAVAO DEL SUR                  67.58711  71.00073  67.06114  78.12500
## COMPOSTELA VALLEY              98.49184  87.14880  93.65245  96.25390
## DAVAO ORIENTAL                 96.25390 112.34226 110.85916 109.59940
## LAGUNA                         38.86719  40.12982  45.88430  40.42483
## BATANGAS                       58.21855  54.71978  53.52743  55.53455
## QUEZON                         35.92368  35.44310  32.89307  37.59698
## CAVITE                         66.18943  60.43284  56.23907  61.39762
## RIZAL                          80.30934 102.04082  86.25818 104.10470
## NEGROS ORIENTAL                55.93388  54.87603  50.43725  49.93141
## CEBU                           25.10886  26.50953  28.19329  25.91418
## BOHOL                          32.34700  30.80926  30.48576  31.86811
## SIQUIJOR                      184.48720 184.48720 186.93552 182.73472
## ILOCOS SUR                     38.23226  42.75729  42.60479  46.81502
## PANGASINAN                     39.52508  39.91527  40.92008  38.62841
## ILOCOS NORTE                   67.75068  67.42019  74.10208  76.09261
## LA UNION                       74.06784  74.88905  71.12196  77.43456
## CAMARINES NORTE                89.88637  99.80832  98.61111 101.01010
## CAMARINES SUR                  31.29520  35.38400  32.50894  33.61144
## ALBAY                          63.61229  57.67184  59.65199  61.00000
## MASBATE                        61.51876  60.52369  60.74249  62.68315
## SORSOGON                       69.44444  79.43755  72.98753  80.78231
## CATANDUANES                   124.17823 120.93174 122.25941 137.05067
## BULACAN                       105.10489 112.40926  87.39076 109.22036
## PAMPANGA                       71.18056  74.85822  66.90279  66.66667
## NUEVA ECIJA                    58.81519  54.43787  51.12099  46.04181
## TARLAC                         76.01330  74.98261  72.40589  76.45429
## ZAMBALES                       79.75230  80.69056  80.38050  80.69056
## AURORA                        141.28728 141.28728 146.11762 169.18153
## BATAAN                         99.48097 100.56228 117.86332 113.22960
## BATANES                       170.13233 226.94628 182.73472 191.13632
## CAGAYAN                        42.83575  43.51182  44.30201  47.23453
## ISABELA                        47.46877  47.04164  49.44116  48.28717
## NUEVA VIZCAYA                  86.71878  73.90320  73.39380  71.13553
## QUIRINO                       160.68240 168.61734 164.64987 164.64987
## ZAMBOANGA DEL NORTE            46.09619  50.83407  45.81404  48.88278
```

```
## ZAMBOANGA SIBUGAY                    68.49544  70.98432  70.68305  71.34364
## ZAMBOANGA DEL SUR                    42.39440  41.54492  41.00987  42.68627
## SOUTHERN LEYTE                       61.87969  66.84743  68.67432  63.33568
## SAMAR                                56.93287  56.52707  59.99409  63.52041
## NORTHERN SAMAR                       63.15507  60.26737  62.62013  60.49149
## EASTERN SAMAR                        57.40995  56.69012  66.81756  69.87311
## LEYTE                                29.97722  29.46303  29.77452  29.56401
## BILIRAN                             150.94795 124.38111 146.11762 175.27816
## MISAMIS ORIENTAL                     49.26483  49.35938  48.22531  52.08333
## BUKIDNON                             49.89408  50.15160  50.94449  57.43802
## CAMIGUIN                            206.93362 203.96438 203.96438 203.96438
## MISAMIS OCCIDENTAL                   73.68421  76.54321  82.23684  81.30612
## LANAO DEL NORTE                      50.23465  62.48205  61.45942  59.02778
## MOUNTAIN PROVINCE                    94.95982  93.42648  91.71333  91.71333
## IFUGAO                               92.20682 104.50106 101.76900  96.30490
## BENGUET                              66.49575  72.10383  72.10383  77.16049
## ABRA                                 58.41424  59.23122  55.34378  54.33062
## APAYAO                              156.98826 188.45467 166.13321 178.32647
## KALINGA                             119.55078 117.13561 126.02613 126.28829
## ROMBLON                              68.67922  77.54330  72.95260  82.82836
## ORIENTAL MINDORO                     70.67719  73.30527  81.28901  76.20845
## PALAWAN                              58.28857  63.04331  57.97770  59.84000
## OCCIDENTAL MINDORO                   95.84242 105.09617 105.09617 124.92564
## MARINDUQUE                          156.71494 180.05912 152.74747 168.61734
## NEGROS OCCIDENTAL                    39.80416  45.88995  41.49499  41.66634
## ANTIQUE                              66.31050  77.85091  74.90134  87.69619
## ILOILO                               30.84126  34.54366  34.58303  35.30024
## CAPIZ                                65.90357  67.79442  69.37218  75.64876
## AKLAN                                76.74162  77.18267  84.55542  85.22727
## GUIMARAS                            196.18382 234.52518 223.05832 246.91358
## SULTAN KUDARAT                       95.36000  97.00176  96.23702  96.23702
## COTABATO                             59.87089  66.48167  64.55643  71.12115
## SARANGANI                           143.70736 134.99782 146.61054 180.05540
## SOUTH COTABATO                       80.61224  80.61224  82.65306  81.91556
## NCR, SECOND DISTRICT                153.08642 158.44875 162.50495 138.87720
## NCR, CITY OF MANILA, FIRST DISTRICT 278.92562 320.24793 309.91736 268.59504
## NCR, FOURTH DISTRICT                106.29130 103.77421 110.38403 116.99385
## NCR, THIRD DISTRICT                 196.18382 185.43402 202.91363 187.30489
## AGUSAN DEL NORTE                     89.74913  97.55463  90.83045 100.48000
## SURIGAO DEL NORTE                    49.99087  59.38354  59.12757  60.69387
## SURIGAO DEL SUR                      72.10859  65.03912  80.43639  79.51183
## AGUSAN DEL SUR                       80.73818 107.74010 100.26298  85.47009
##                                          2016
## LANAO DEL SUR                        67.80187
## MAGUINDANAO                         102.51912
## SULU                                 96.14512
## BASILAN                             121.16571
## TAWI-TAWI                           153.89351
## DAVAO DEL NORTE                      91.82736
## DAVAO DEL SUR                        84.87654
## COMPOSTELA VALLEY                   106.65973
## DAVAO ORIENTAL                      113.37868
## LAGUNA                               40.61664
## BATANGAS                             51.71658
```

```
## QUEZON                        35.60091
## CAVITE                        60.86542
## RIZAL                        103.36109
## NEGROS ORIENTAL               55.90257
## CEBU                          25.71229
## BOHOL                         32.42414
## SIQUIJOR                     199.53791
## ILOCOS SUR                    43.06982
## PANGASINAN                    42.88000
## ILOCOS NORTE                  78.73738
## LA UNION                      79.36508
## CAMARINES NORTE               95.78971
## CAMARINES SUR                 33.73366
## ALBAY                         65.40239
## MASBATE                       60.74249
## SORSOGON                      77.94785
## CATANDUANES                  107.23311
## BULACAN                      107.32323
## PAMPANGA                      68.64285
## NUEVA ECIJA                   53.60536
## TARLAC                        71.37520
## ZAMBALES                      76.42998
## AURORA                       138.87212
## BATAAN                       115.77503
## BATANES                      191.13632
## CAGAYAN                       47.43633
## ISABELA                       48.56053
## NUEVA VIZCAYA                 80.16862
## QUIRINO                      168.61734
## ZAMBOANGA DEL NORTE           47.55445
## ZAMBOANGA SIBUGAY             71.16843
## ZAMBOANGA DEL SUR             43.40278
## SOUTHERN LEYTE                64.30634
## SAMAR                         60.76489
## NORTHERN SAMAR                58.28012
## EASTERN SAMAR                 66.93236
## LEYTE                         30.34693
## BILIRAN                      170.26929
## MISAMIS ORIENTAL              53.04784
## BUKIDNON                      53.50773
## CAMIGUIN                     238.43723
## MISAMIS OCCIDENTAL            84.76454
## LANAO DEL NORTE               57.57972
## MOUNTAIN PROVINCE             93.33658
## IFUGAO                       101.76900
## BENGUET                       68.09806
## ABRA                          56.35694
## APAYAO                       172.22984
## KALINGA                      119.08891
## ROMBLON                       68.06874
## ORIENTAL MINDORO              85.60185
## PALAWAN                       59.74415
## OCCIDENTAL MINDORO           127.56957
## MARINDUQUE                   152.74747
```

```
## NEGROS OCCIDENTAL                       41.20123
## ANTIQUE                                 76.47817
## ILOILO                                  35.02155
## CAPIZ                                   79.19991
## AKLAN                                   86.83837
## GUIMARAS                               244.55791
## SULTAN KUDARAT                         114.55059
## COTABATO                                67.25244
## SARANGANI                              197.03509
## SOUTH COTABATO                          78.57143
## NCR, SECOND DISTRICT                   138.87720
## NCR, CITY OF MANILA, FIRST DISTRICT 265.00811
## NCR, FOURTH DISTRICT                   123.60368
## NCR, THIRD DISTRICT                    213.31946
## AGUSAN DEL NORTE                        96.23702
## SURIGAO DEL NORTE                       66.44376
## SURIGAO DEL SUR                         73.50222
## AGUSAN DEL SUR                          89.57922
```

Save as CSV

```
write.csv(hhi_df, "unweighted_hhi_yearly.csv")
```

HHI PH

Rafael Ma. Vicente Tagulao

2024-12-05

```r
library(sf)
```

```
## Linking to GEOS 3.12.1, GDAL 3.8.4, PROJ 9.3.1; sf_use_s2() is TRUE
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(stringdist)
library(tidyr)
```

```
##
## Attaching package: 'tidyr'

## The following object is masked from 'package:stringdist':
##
##     extract
```

```r
library(ggplot2)
```

```r
#Reading the calculated HHI values
hhi = read.csv("datasets/unweighted_hhi_yearly.csv")
```

```r
# Reading shapefile obtained from GitHub
ph <- st_read("datasets/raw_geo_dataset/Provdists.shp.shp")
```

```
## Reading layer 'Provdists.shp' from data source
##   'C:\Users\asus\OneDrive\Documents\Undergraduate\Y4S1\ECON 185.78i\Project\datasets\raw_geo_dataset
##   using driver 'ESRI Shapefile'
## Simple feature collection with 88 features and 18 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 114.2779 ymin: 4.587294 xmax: 126.605 ymax: 21.12189
## Geodetic CRS:  WGS 84
```

```
# Merging Maguindanao del Norte and Maguindanao del Sur into Maguindanao to fit HHI values
ph <- ph %>%
  mutate(name = ifelse(name %in% c("Maguindanao del Norte", "Maguindanao del Sur"),
                       "Maguindanao", name)) %>%
  group_by(name) %>%
  summarize(geometry = st_combine(geometry), .groups = "drop")

# Ensuring that the geometry has no intersecting areas
ph <- ph %>%
  mutate(geometry = st_make_valid(geometry))

# Calculating the centroid of polygons and extracting the latitude and longitude
ph <- ph %>%
  mutate(
    centroid = st_centroid(geometry),
    lon = st_coordinates(centroid)[, 1],
    lat = st_coordinates(centroid)[, 2]
  )
```

```
# Changing the columns to lower case, to match them better, and removing a null row in the shapefile
ph <- ph %>% mutate(name_lower = tolower(name))
hhi <- hhi %>% mutate(X_lower = tolower(X))
ph <- ph %>%
  filter(!is.na(name_lower))

# Manually assigning matching
hhi <- hhi %>%
  mutate(
    X_lower = case_when(
      X_lower == "ncr, city of manila, first district" ~ "ncr, city of manila, first district (not a pr
      X_lower == "ncr, second district" ~ "ncr, second district (not a province)",
      X_lower == "ncr, third district" ~ "ncr, third district (not a province)",
      X_lower == "ncr, fourth district" ~ "ncr, fourth district (not a province)",
      X_lower == "compostela valley" ~ "davao de oro",
      TRUE ~ X_lower
    )
  )

hhi_map <- hhi %>%
  left_join(ph, by = c("X_lower" = "name_lower"))

#This is the the object we will then use for the visualization
hhi_map <- hhi_map %>%
  select(X, X2004, X2007, X2010, X2013, X2016, geometry, lon, lat)
```

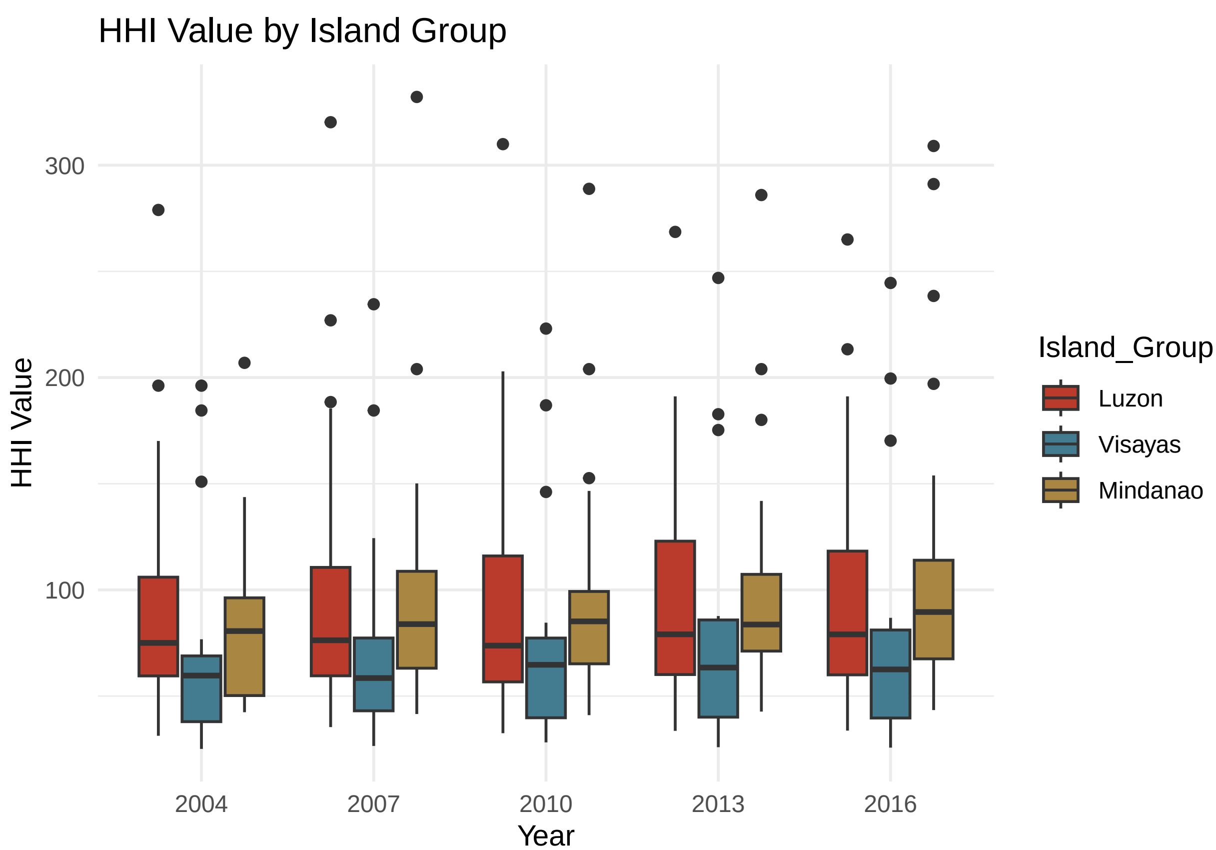
```r
# Exporting it now to shapefile
st_write(hhi_map, "hhi_map.shp", delete_dsn = TRUE)
```

```
## Warning in CPL_write_ogr(obj, dsn, layer, driver,
## as.character(dataset_options), : GDAL Error 1: hhi_map.shp does not appear to
## be a file or directory.

## Deleting source 'hhi_map.shp' failed
## Writing layer 'hhi_map' to data source 'hhi_map.shp' using driver 'ESRI Shapefile'
## Writing 85 features with 8 fields and geometry type Multi Polygon.
```

```r
# Now, adding island group names

hhi_map <- hhi_map %>%
  mutate(
    Island_Group = case_when(
      X %in% c("ABRA", "ALBAY", "APAYAO", "AURORA", "BATAAN", "BATANES", "BATANGAS",
               "BENGUET", "BULACAN", "CAGAYAN", "CAMARINES NORTE", "CAMARINES SUR",
               "CATANDUANES", "CAVITE", "IFUGAO", "ILOCOS NORTE", "ILOCOS SUR",
               "ISABELA", "KALINGA", "LA UNION", "LAGUNA", "MARINDUQUE", "MASBATE",
               "MOUNTAIN PROVINCE", "NUEVA ECIJA", "NUEVA VIZCAYA", "OCCIDENTAL MINDORO",
               "ORIENTAL MINDORO", "PALAWAN", "PAMPANGA", "PANGASINAN", "QUEZON",
               "QUIRINO", "RIZAL", "ROMBLON", "SORSOGON", "TARLAC", "ZAMBALES",
               "NCR, SECOND DISTRICT", "NCR, CITY OF MANILA, FIRST DISTRICT",
               "NCR, FOURTH DISTRICT", "NCR, THIRD DISTRICT") ~ "Luzon",
      X %in% c("AKLAN", "ANTIQUE", "BILIRAN", "BOHOL", "CAPIZ", "CEBU", "EASTERN SAMAR",
               "GUIMARAS", "ILOILO", "LEYTE", "NEGROS OCCIDENTAL", "NEGROS ORIENTAL",
               "NORTHERN SAMAR", "SAMAR", "SIQUIJOR", "SOUTHERN LEYTE") ~ "Visayas",
      X %in% c("AGUSAN DEL NORTE", "AGUSAN DEL SUR", "BASILAN", "BUKIDNON",
               "CAMIGUIN", "COMPOSTELA VALLEY", "COTABATO", "DAVAO DE ORO", "DAVAO DEL NORTE", "DAVAO DE
               "DAVAO OCCIDENTAL", "DAVAO ORIENTAL", "DINAGAT ISLANDS", "LANAO DEL NORTE",
               "LANAO DEL SUR", "MAGUINDANAO",
               "MISAMIS OCCIDENTAL", "MISAMIS ORIENTAL", "SARANGANI", "SOUTH COTABATO",
               "SULTAN KUDARAT", "SULU", "SURIGAO DEL NORTE", "SURIGAO DEL SUR",
               "TAWI-TAWI", "ZAMBOANGA DEL NORTE", "ZAMBOANGA DEL SUR",
               "ZAMBOANGA SIBUGAY") ~ "Mindanao",
      TRUE ~ NA_character_
    )
  )
```

```r
# Converting the dataframe into long
names(hhi_map)[-1] <- gsub("^X", "", names(hhi_map)[-1])
data_long <- hhi_map %>%
  pivot_longer(
    cols = c("2004", "2007", "2010", "2013", "2016"),
    names_to = "Year",
    values_to = "Value"
  ) %>%
  mutate(
    Year = as.numeric(Year),
    Island_Group = factor(Island_Group, levels = c("Luzon", "Visayas", "Mindanao"))
  ) %>%
  filter(Value != 0)
```

```r
# Box Plot
ggplot(data_long, aes(x = as.factor(Year), y = Value, fill = Island_Group)) +
  geom_boxplot() +
  labs(title = "HHI Value by Island Group",
       x = "Year", y = "HHI Value") +
  theme_minimal() +
  scale_fill_manual(
    values = c(
      "Luzon" = "#BA3A2C",
      "Visayas" = "#437C90",
      "Mindanao" = "#A98743"
    )
  )
```



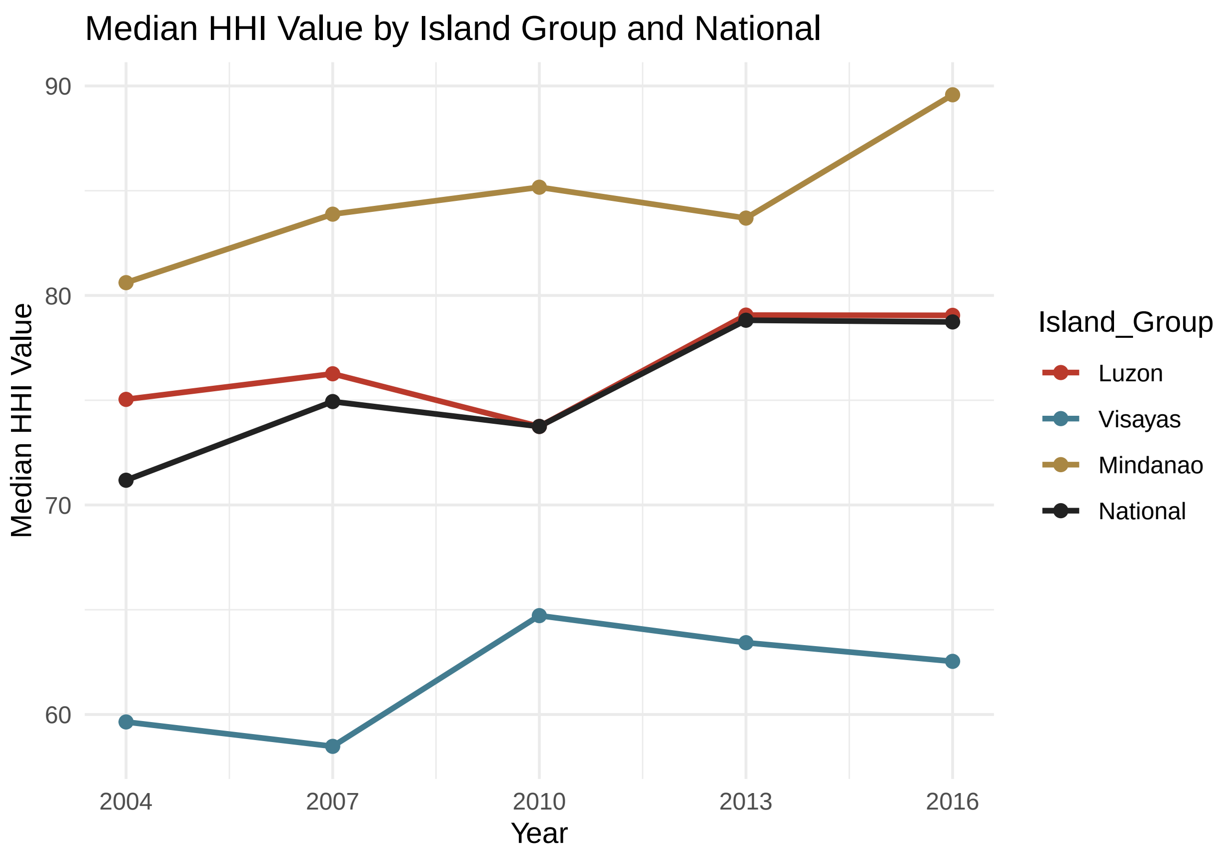

```r
# Computing for median of each group
median_data <- data_long %>%
  group_by(Island_Group, Year) %>%
  summarize(Median_Value = median(Value, na.rm = TRUE)) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'Island_Group'. You can override using the
## `.groups` argument.
```

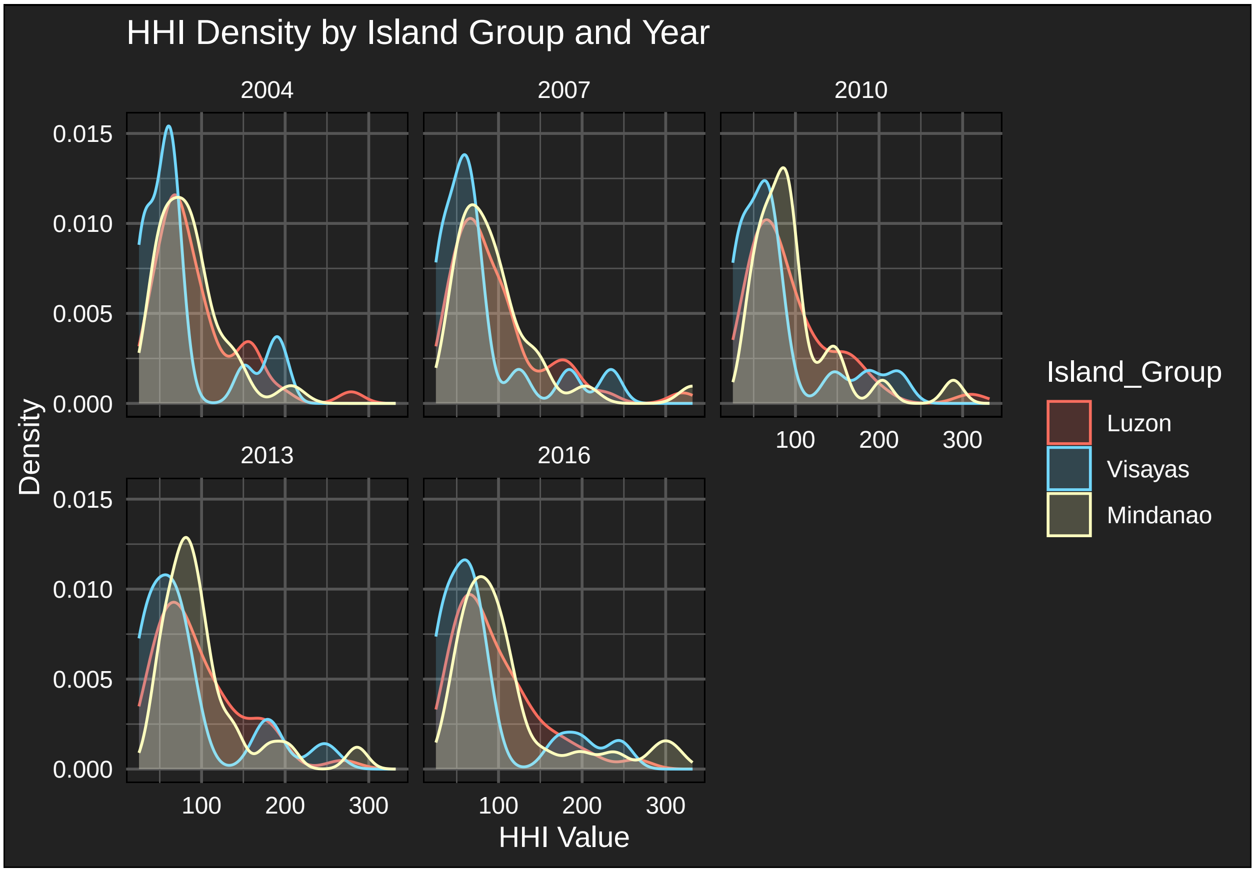
```r
# Computing  the national median
national_median <- data_long %>%
  group_by(Year) %>%
  summarize(Median_Value = median(Value, na.rm = TRUE)) %>%
  mutate(Island_Group = "National")

combined_data <- bind_rows(median_data, national_median)
```

```r
combined_data$Island_Group <- factor(combined_data$Island_Group, levels = c("Luzon", "Visayas", "Mindana

# LINE PLOT
ggplot(combined_data, aes(x = Year, y = Median_Value, color = Island_Group, group = Island_Group)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  labs(title = "Median HHI Value by Island Group and National",
       x = "Year", y = "Median HHI Value") +
  scale_color_manual(
    values = c(
      "Luzon" = "#BA3A2C",
      "Visayas" = "#437C90",
      "Mindanao" = "#A98743",
      "National" = "#222222"
    )
  ) +
  theme_minimal()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

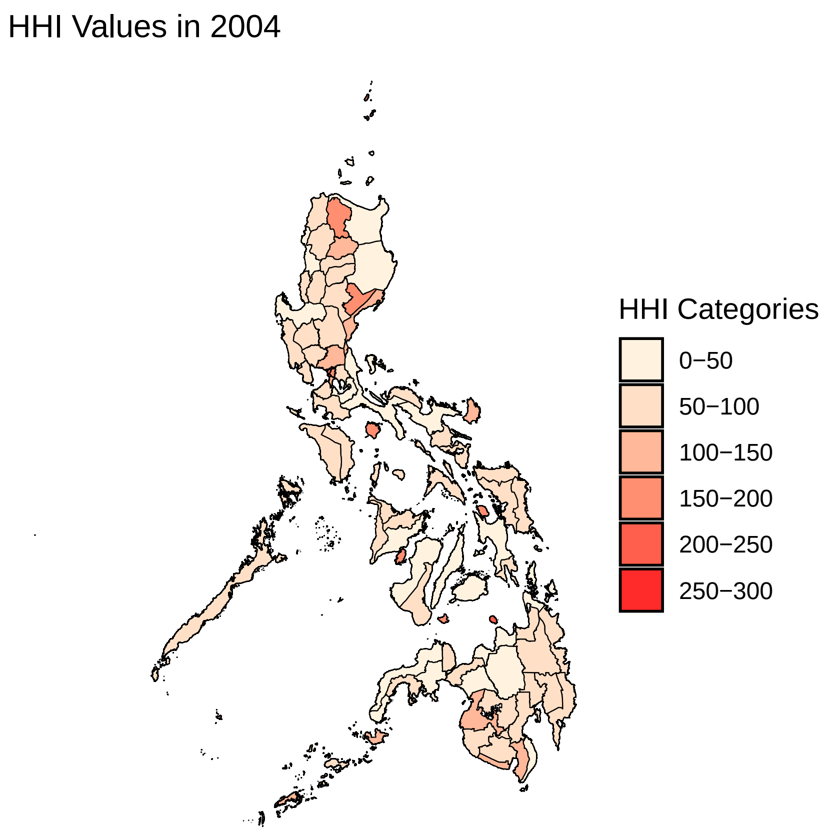



```r
# Colors for the island groups
cols <- c("Luzon" = "#F76D5E", "Visayas" = "#72D8FF", "Mindanao" = "#FFFFBF")

# DENSITY PLOT
ggplot(data_long, aes(x = Value, fill = Island_Group, color = Island_Group)) +
  geom_density(alpha = 0.2) +
  scale_fill_manual(values = cols) +
  scale_color_manual(values = cols) +
  facet_wrap(~ Year) +
  labs(title = "HHI Density by Island Group and Year",
       x = "HHI Value", y = "Density") +
  theme_minimal() +
  theme(
    plot.background = element_rect(fill = "#222222"),
    panel.background = element_rect(fill = "#222222"),
    panel.grid.major = element_line(color = "#555555"),
    panel.grid.minor = element_line(color = "#555555"),
    axis.text = element_text(color = "white"),
    axis.title = element_text(color = "white"),
    plot.title = element_text(color = "white"),
    legend.text = element_text(color = "white"),
    legend.title = element_text(color = "white"),
    strip.text = element_text(color = "white")
  )
```



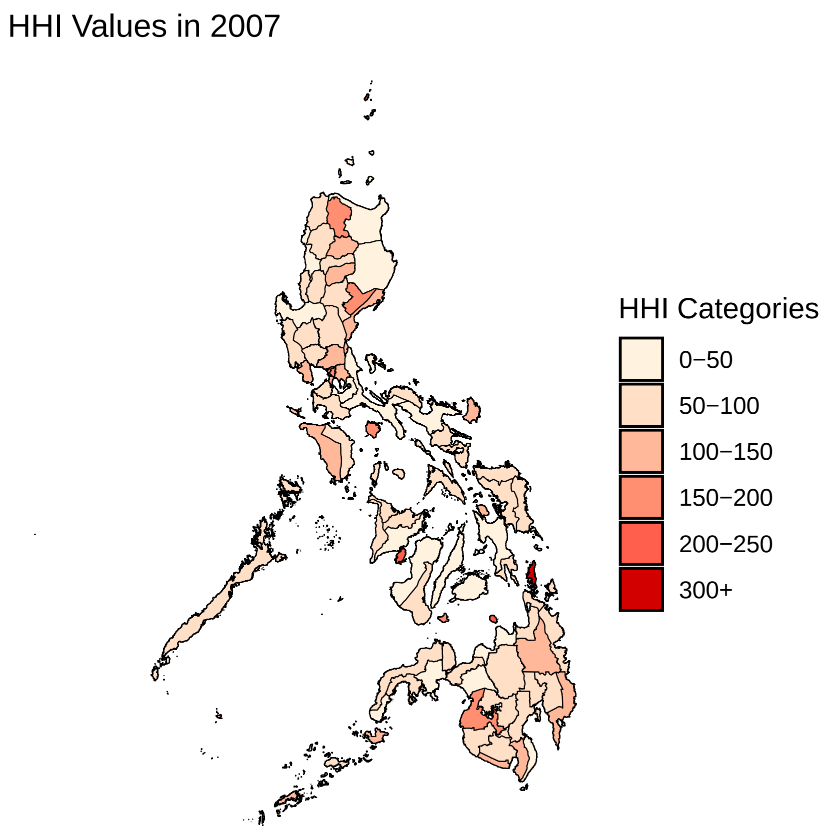

```r
# PLOTTING IN MAP
#Year 2004
hhi_map <- st_as_sf(hhi_map)

# Create bins for HHI values
hhi_map$hhi_bins <- cut(
  hhi_map$`2004`,
  breaks = c(0, 50, 100, 150, 200, 250, 300, Inf),
  labels = c("0-50", "50-100", "100-150", "150-200", "200-250", "250-300", "300+"),
  include.lowest = TRUE
)

category_colors <- c(
  "0-50" = "#FFF3E0",
  "50-100" = "#FFE0C7",
  "100-150" = "#FFB899",
  "150-200" = "#FF8F70",
  "200-250" = "#FF5F4C",
  "250-300" = "#FF2A2A",
  "300+" = "#D30000"
)

plot1 <- ggplot(hhi_map) +
  geom_sf(aes(fill = hhi_bins), color = "black", alpha = 1) +
  scale_fill_manual(values = category_colors, na.value = "grey80", name = "HHI Categories") +
  theme_void() +
```

```r
  ggtitle("HHI Values in 2004") +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 12)
  )

plot1
```

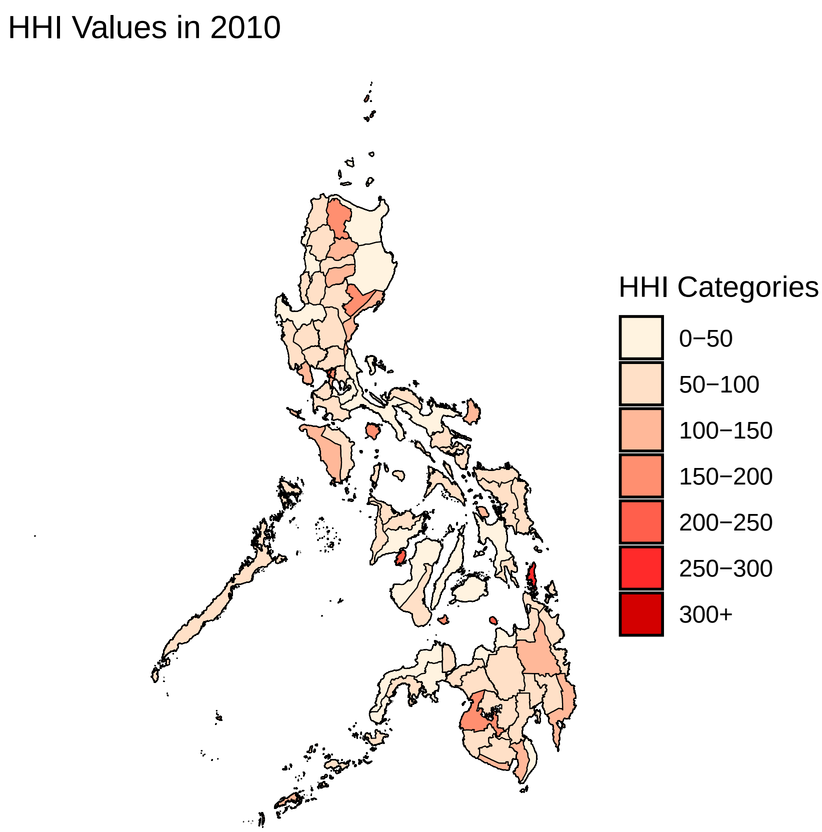



```r
#Year 2007
hhi_map$hhi_bins <- cut(
  hhi_map$`2007`,
  breaks = c(0, 50, 100, 150, 200, 250, 300, Inf),
  labels = c("0-50", "50-100", "100-150", "150-200", "200-250", "250-300", "300+"),
  include.lowest = TRUE
)

plot2 <- ggplot(hhi_map) +
  geom_sf(aes(fill = hhi_bins), color = "black", alpha = 1) +
  scale_fill_manual(values = category_colors, na.value = "grey80", name = "HHI Categories") +
  theme_void() +
  ggtitle("HHI Values in 2007") +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 12)
  )
```

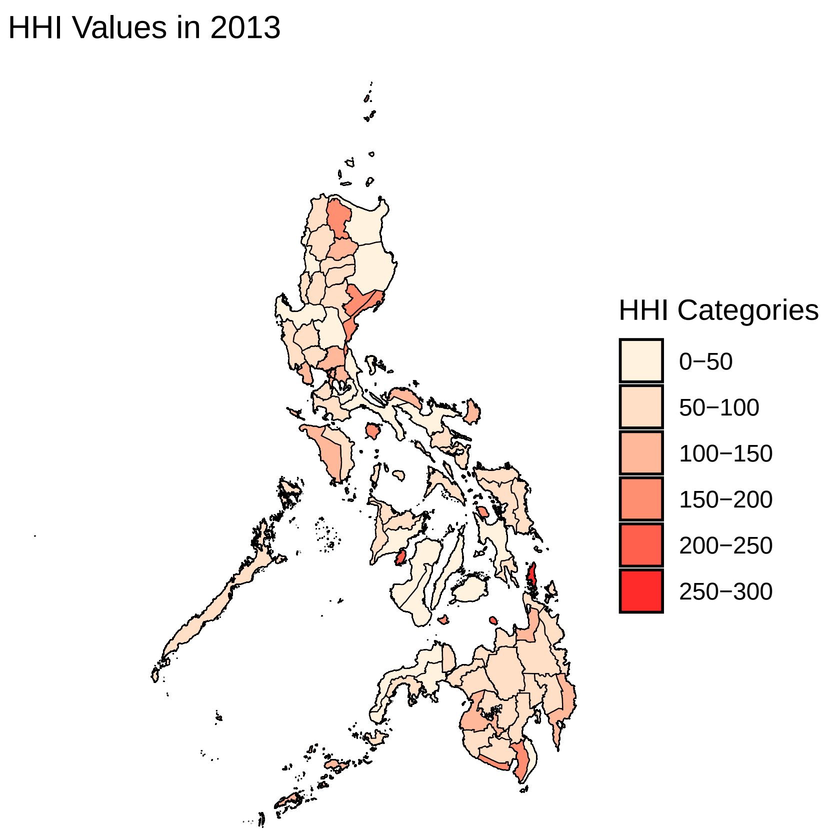
```r
plot2
```




```r
#Year 2010
hhi_map$hhi_bins <- cut(
  hhi_map$`2010`,
  breaks = c(0, 50, 100, 150, 200, 250, 300, Inf),
  labels = c("0-50", "50-100", "100-150", "150-200", "200-250", "250-300", "300+"),
  include.lowest = TRUE
)

plot3 <- ggplot(hhi_map) +
  geom_sf(aes(fill = hhi_bins), color = "black", alpha = 1) +
  scale_fill_manual(values = category_colors, na.value = "grey80", name = "HHI Categories") +
  theme_void() +
  ggtitle("HHI Values in 2010") +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 12)
  )

plot3
```

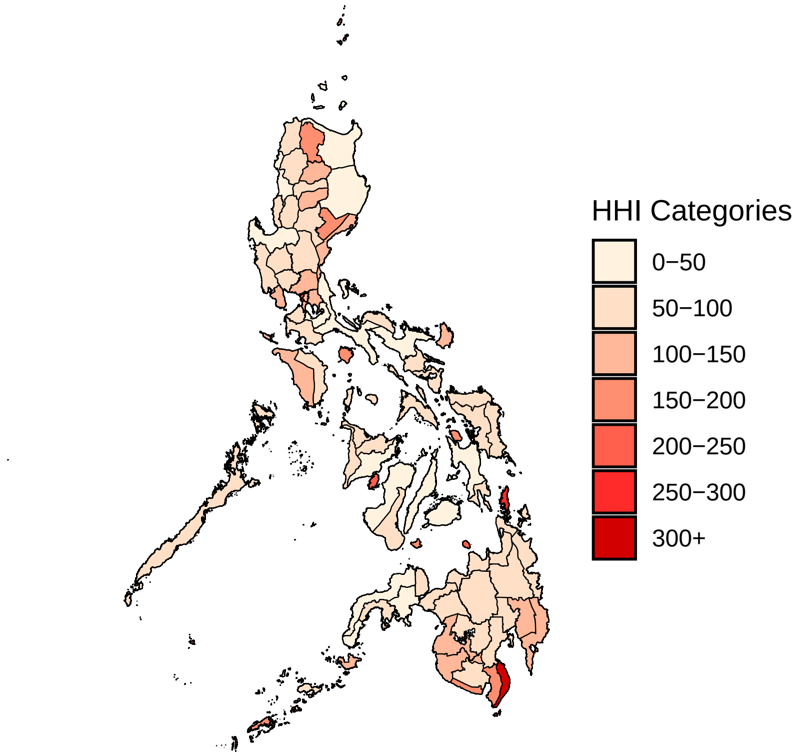



```r
#Year 2013
hhi_map$hhi_bins <- cut(
  hhi_map$`2013`,
  breaks = c(0, 50, 100, 150, 200, 250, 300, Inf),
  labels = c("0-50", "50-100", "100-150", "150-200", "200-250", "250-300", "300+"),
  include.lowest = TRUE
)

plot4 <- ggplot(hhi_map) +
  geom_sf(aes(fill = hhi_bins), color = "black", alpha = 1) +
  scale_fill_manual(values = category_colors, na.value = "grey80", name = "HHI Categories") +
  theme_void() +
  ggtitle("HHI Values in 2013") +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 12)
  )

plot4
```




```r
#Year 2016
hhi_map$hhi_bins <- cut(
  hhi_map$`2016`,
  breaks = c(0, 50, 100, 150, 200, 250, 300, Inf),
  labels = c("0-50", "50-100", "100-150", "150-200", "200-250", "250-300", "300+"),
  include.lowest = TRUE
)

plot5 <- ggplot(hhi_map) +
  geom_sf(aes(fill = hhi_bins), color = "black", alpha = 1) +
  scale_fill_manual(values = category_colors, na.value = "grey80", name = "HHI Categories") +
  theme_void() +
  ggtitle("HHI Values in 2016") +
  theme(
    legend.position = "right",
    plot.title = element_text(size = 12)
  )

plot5
```

HHI Values in 2016



FIES Data Cleaning

Sted Micah Cheng

2024-12-09

```r
# libraries needed
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(sf)
```

```
## Linking to GEOS 3.12.1, GDAL 3.8.4, PROJ 9.3.1; sf_use_s2() is TRUE
```

Import data

```r
fies_2006 = read.csv('datasets/fies_datasets/FIES PUF 2006 Vol.1.csv')
fies_2009 = read.csv('datasets/fies_datasets/FIES PUF 2009 Vol.2.csv')
fies_2012 = read.csv('datasets/fies_datasets/FIES PUF 2012 Vol.1.csv')
fies_2015 = read.csv('datasets/fies_datasets/FIES PUF 2015 Vol.1.csv')
fies_2018 = read.csv('datasets/fies_datasets/FIES PUF 2018 Vol.1.csv')
hhi_map = st_read('datasets/hhi_map/hhi_map.shp')
```

```
## Reading layer `hhi_map' from data source
##   `C:\Users\asus\OneDrive\Documents\Undergraduate\Y4S1\ECON 185.78i\Project\datasets\hhi_map\hhi_map
##   using driver `ESRI Shapefile'
## Simple feature collection with 85 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 114.2779 ymin: 4.587294 xmax: 126.605 ymax: 21.12189
## Geodetic CRS:  WGS 84
```

Select the relevant column names

```r
col_names_2006 = c('W_HCN', 'FSIZE', 'WAGES', 'NETSH', 'CONAB', 'CONDO', 'RENTL', 'INTRS',
                   'PNSNS', 'DVDND', 'OTHIN', 'IFAMS', 'REGFT', 'EACFG', 'EALPR', 'EAFIS',
                   'EAFOR', 'EATRD', 'EAMFG', 'EACPS', 'EATCS', 'EAMNG', 'EACON', 'EANEC', 'EAINC',
                   'TOINC', 'CREAL', 'MEAT', 'FISHM', 'DAIRY', 'FRUIT', 'FDNEC', 'COFCT', 'NONAL',
                   'ALBEV', 'TBCCO', 'FHOME', 'FDOUT', 'FOOD', 'CLOTH', 'NDFUR', 'MEDIC',
                   'FUEL', 'TRCOM', 'RCRTN', 'EDUC', 'OTHEX',
                   'OTDIS', 'NONFOOD', 'TOEXP', 'TOTDIS', 'OTHRE', 'TOREC')

col_names_2009 = c('W_ID', 'FSIZE', 'WAGES', 'NETSH', 'CONAB', 'CONDO', 'RENTL', 'INTRS',
                   'PNSNS', 'DVDND', 'OSINC', 'IFAMS', 'REGFT', 'EACFG', 'EALPR', 'EAFIS',
                   'EAFOR', 'EATRD', 'EAMFG', 'EACPS', 'EATCS', 'EAMNG', 'EACON', 'EANEC', 'EAINC',
                   'TOINC', 'CREAL', 'MEAT', 'FISHM', 'TMILK', 'FRUIT', 'FDNEC', 'COFCT', 'NONAL',
                   'ALBEV', 'TBCCO', 'FHOME', 'FDOUT', 'FOOD', 'CLOTH', 'NDFUR', 'MEDIC',
                   'FUEL', 'TRCOM', 'RCRTN', 'EDUC', 'OTHEX',
                   'OTDIS', 'NFOOD', 'TOTEX', 'TOTDI', 'OTREC', 'TOREC')

col_names_2012 = c('W_OID', 'FSIZE', 'WAGES', 'NETSHARE', 'CASH_ABROAD', 'CASH_DOMESTIC', 'RENTALS_REC'
                   'PENSION', 'DIVIDENDS', 'OTHER_SOURCE', 'NET_RECEIPT', 'REGFT', 'NET_CFG', 'NET_LPR'
                   'NET_FOR', 'NET_RET', 'NET_MFG', 'NET_COM', 'NET_TRANS', 'NET_MIN', 'NET_CONS', 'NET_
                   'TOINC', 'T_BREAD', 'T_MEAT', 'T_FISH', 'T_MILK', 'T_FRUIT', 'T_VEG', 'T_FOOD_NEC',
                   'T_ALCOHOL', 'T_TOBACCO', 'T_FOOD_HOME', 'T_FOOD_OUTSIDE', 'T_FOOD', 'T_CLOTH', 'T_FU
                   'T_HOUSING_WATER', 'T_TRANSPORT', 'T_COMMUNICATION', 'T_RECREATION', 'T_EDUCATION',
                   'T_OTHER_DISBURSEMENT', 'T_NFOOD', 'T_TOTEX', 'T_TOTDIS', 'T_OTHREC', 'T_TOREC')

col_names_2015 = c('W_ID', 'FSIZE', 'WAGES', 'NETSHARE', 'CASH_ABROAD', 'CASH_DOMESTIC', 'RENTALS_REC',
                   'PENSION', 'DIVIDENDS', 'OTHER_SOURCE', 'NET_RECEIPT', 'REGFT', 'NET_CFG', 'NET_LPR'
                   'NET_FOR', 'NET_RET', 'NET_MFG', 'NET_COM', 'NET_TRANS', 'NET_MIN', 'NET_CONS', 'NET_
                   'TOINC', 'BREAD', 'MEAT', 'FISH', 'MILK', 'FRUIT', 'VEG', 'FOOD_NEC', 'COFFEE', 'MINE
                   'ALCOHOL', 'TOBACCO', 'FOOD_HOME', 'FOOD_OUTSIDE', 'FOOD', 'CLOTH', 'FURNISHING', 'HE
                   'HOUSING_WATER', 'TRANSPORT', 'COMMUNICATION', 'RECREATION', 'EDUCATION', 'OTHER_EXPE
                   'OTHER_DISBURSEMENT', 'NFOOD', 'TOTEX', 'TOTDIS', 'OTHREC', 'TOREC')

col_names_2018 = c('W_PROV', 'FSIZE', 'WAGES', 'NETSHARE', 'CASH_ABROAD', 'CASH_DOMESTIC', 'RENTALS_REC
                   'PENSION', 'DIVIDENDS', 'OTHER_SOURCE', 'NET_RECEIPT', 'REGFT', 'NET_CFG', 'NET_LPR'
                   'NET_FOR', 'NET_RET', 'NET_MFG', 'NET_COM', 'NET_TRANS', 'NET_MIN', 'NET_CONS', 'NET_
                   'TOINC', 'BREAD', 'MEAT', 'FISH', 'MILK', 'FRUIT', 'VEG', 'FOOD_NEC', 'COFFEE', 'MINE
                   'ALCOHOL', 'TOBACCO', 'FOOD_HOME', 'FOOD_OUTSIDE', 'FOOD', 'CLOTH', 'FURNISHING', 'HE
                   'HOUSING_WATER', 'TRANSPORT', 'COMMUNICATION', 'RECREATION', 'EDUCATION', 'OTHER_EXPE
                   'OTHER_DISBURSEMENT', 'NFOOD', 'TOTEX', 'TOTDIS', 'OTHREC', 'TOREC')

fies_2006_selected = fies_2006 %>% select(all_of(col_names_2006))
fies_2009_selected = fies_2009 %>% select(all_of(col_names_2009))
fies_2012_selected = fies_2012 %>% select(all_of(col_names_2012))
fies_2015_selected = fies_2015 %>% select(all_of(col_names_2015))
fies_2018_selected = fies_2018 %>% select(all_of(col_names_2018))
```

Extract the province number using the ID

```r
fies_2006_selected = fies_2006_selected %>% mutate(W_PROV = W_HCN %/% 10^16)
fies_2009_selected = fies_2009_selected %>% mutate(W_PROV = W_ID %/% 10^8)
fies_2012_selected = fies_2012_selected %>% mutate(W_PROV = W_OID %/% 10^8)
fies_2015_selected = fies_2015_selected %>% mutate(W_PROV = W_ID %/% 10^8)
```

Group by province and take the mean per column

```r
province_means = function(df_selected) {
  df_aggregated = df_selected %>%
    group_by(W_PROV) %>%
    summarize(across(everything(), \(x) mean(x, na.rm = TRUE)))
  return(df_aggregated)
}

fies_2006_aggregated = province_means(fies_2006_selected)
fies_2009_aggregated = province_means(fies_2009_selected)
fies_2012_aggregated = province_means(fies_2012_selected)
fies_2015_aggregated = province_means(fies_2015_selected)
fies_2018_aggregated = province_means(fies_2018_selected)
```

Remove helper ID variable and provinces not in all FIES datasets

```r
# remove the helper ID variable
fies_2006_aggregated = fies_2006_aggregated %>% select(-W_HCN)
fies_2009_aggregated = fies_2009_aggregated %>% select(-W_ID)
fies_2012_aggregated = fies_2012_aggregated %>% select(-W_OID)
fies_2015_aggregated = fies_2015_aggregated %>% select(-W_ID)

# remove provinces that are in 2018 but not in the rest of the datasets
fies_2018_aggregated = fies_2018_aggregated %>% filter(W_PROV != 85 & W_PROV != 86)
```

Create new columns

```r
# 2006 and 2009 have a combined column for fruits and vegetables,
# and another combined column for transportation and communication
# while 2012, 2015, and 2018 have two separate columns
fies_2012_aggregated = fies_2012_aggregated %>%
  mutate(FRUITVEG = T_FRUIT + T_VEG) %>%
  mutate(TRCOM = T_TRANSPORT + T_COMMUNICATION) %>%
  select(-T_FRUIT, -T_VEG, -T_TRANSPORT, -T_COMMUNICATION)
fies_2015_aggregated = fies_2015_aggregated %>%
  mutate(FRUITVEG = FRUIT + VEG) %>%
  mutate(TRCOM = TRANSPORT + COMMUNICATION) %>%
  select(-FRUIT, -VEG, -TRANSPORT, -COMMUNICATION)
fies_2018_aggregated = fies_2018_aggregated %>%
  mutate(FRUITVEG = FRUIT + VEG) %>%
  mutate(TRCOM = TRANSPORT + COMMUNICATION) %>%
  select(-FRUIT, -VEG, -TRANSPORT, -COMMUNICATION)
```

Add column that signifies the year for the unique idenitifier later

```r
fies_2006_aggregated = fies_2006_aggregated %>% mutate(YEAR = 2006)
fies_2009_aggregated = fies_2009_aggregated %>% mutate(YEAR = 2009)
fies_2012_aggregated = fies_2012_aggregated %>% mutate(YEAR = 2012)
fies_2015_aggregated = fies_2015_aggregated %>% mutate(YEAR = 2015)
fies_2018_aggregated = fies_2018_aggregated %>% mutate(YEAR = 2018)
```

Make all five datasets have the same column names (follow the 2018 column names)

```r
colnames(fies_2006_aggregated) = colnames(fies_2018_aggregated)
colnames(fies_2009_aggregated) = colnames(fies_2018_aggregated)
colnames(fies_2012_aggregated) = colnames(fies_2018_aggregated)
colnames(fies_2015_aggregated) = colnames(fies_2018_aggregated)
```

Replace the provincial codes with the actual province names

```r
# AI was used to convert the information in the metadata into this list
replacement_dict = list(
  "1" = "ABRA",
  "2" = "AGUSAN DEL NORTE", # manual correction made
  "3" = "AGUSAN DEL SUR",
  "4" = "AKLAN",
  "5" = "ALBAY",
  "6" = "ANTIQUE",
  "7" = "BASILAN",
  "8" = "BATAAN",
  "9" = "BATANES",
  "10" = "BATANGAS",
  "11" = "BENGUET",
  "12" = "BOHOL",
  "13" = "BUKIDNON",
  "14" = "BULACAN",
  "15" = "CAGAYAN",
  "16" = "CAMARINES NORTE",
  "17" = "CAMARINES SUR",
  "18" = "CAMIGUIN",
  "19" = "CAPIZ",
  "20" = "CATANDUANES",
  "21" = "CAVITE",
  "22" = "CEBU",
  "23" = "DAVAO",
  "24" = "DAVAO DEL SUR", # manual correction made
  "25" = "DAVAO ORIENTAL",
  "26" = "EASTERN SAMAR",
  "27" = "IFUGAO",
  "28" = "ILOCOS NORTE",
  "29" = "ILOCOS SUR",
  "30" = "ILOILO",
  "31" = "ISABELA",
  "32" = "KALINGA",
  "33" = "LA UNION",
  "34" = "LAGUNA",
  "35" = "LANAO DEL NORTE",
  "36" = "LANAO DEL SUR",
  "37" = "LEYTE",
  "38" = "MAGUINDANAO",
  "39" = "NCR, CITY OF MANILA, FIRST DISTRICT", # manual correction made
  "40" = "MARINDUQUE",
  "41" = "MASBATE",
  "42" = "MISAMIS OCCIDENTAL",
  "43" = "MISAMIS ORIENTAL",
  "44" = "MOUNTAIN PROVINCE",
```

```
  "45" = "NEGROS OCCIDENTAL",
  "46" = "NEGROS ORIENTAL",
  "47" = "COTABATO",
  "48" = "NORTHERN SAMAR",
  "49" = "NUEVA ECIJA",
  "50" = "NUEVA VIZCAYA",
  "51" = "OCCIDENTAL MINDORO",
  "52" = "ORIENTAL MINDORO",
  "53" = "PALAWAN",
  "54" = "PAMPANGA",
  "55" = "PANGASINAN",
  "56" = "QUEZON",
  "57" = "QUIRINO",
  "58" = "RIZAL",
  "59" = "ROMBLON",
  "60" = "SAMAR", # manual correction made
  "61" = "SIQUIJOR",
  "62" = "SORSOGON",
  "63" = "SOUTH COTABATO",
  "64" = "SOUTHERN LEYTE",
  "65" = "SULTAN KUDARAT",
  "66" = "SULU",
  "67" = "SURIGAO DEL NORTE",
  "68" = "SURIGAO DEL SUR",
  "69" = "TARLAC",
  "70" = "TAWI-TAWI",
  "71" = "ZAMBALES",
  "72" = "ZAMBOANGA DEL NORTE",
  "73" = "ZAMBOANGA DEL SUR",
  "74" = "NCR, SECOND DISTRICT", # manual correction made
  "75" = "NCR, THIRD DISTRICT", # manual correction made
  "76" = "NCR, FOURTH DISTRICT", # manual correction made
  "77" = "AURORA",
  "78" = "BILIRAN",
  "79" = "GUIMARAS",
  "80" = "SARANGANI",
  "81" = "APAYAO",
  "82" = "COMPOSTELA VALLEY",
  "83" = "ZAMBOANGA SIBUGAY"
)
```

Remove provinces that are only in one dataset; create unique identifier with province and year

```
# Davao Del Norte, Davao Occidental, and Dinagat Islands are in HHI but not in FIES
hhi_map = hhi_map %>%
  arrange(X) %>%
  filter(X != "DAVAO DEL NORTE" & X != "DAVAO OCCIDENTAL" & X != "DINAGAT ISLANDS")

# create a unique identifier for each row in the FIES datasets
add_prov_year_identifier = function(df_aggregated) {
  df_aggregated = df_aggregated %>%

    # 23 (Davao), 97 (Isabela City), and 98 (Cotabato City) are in FIES but not in HHI
    filter(W_PROV != 23 & W_PROV != 97 & W_PROV != 98) %>%
```

```
    # apply the replacement dictionary
    mutate(W_PROV = recode(W_PROV, !!!replacement_dict)) %>%

    # make the W_PROV column contain both the province and the year so that there is only one identifie
    mutate(W_PROV = paste(W_PROV, YEAR, sep = "_")) %>%

    # remove the year
    select(-YEAR) %>%

    # rename W_PROV to PROV_YEAR
    rename(PROV_YEAR = W_PROV) %>%

    # arrange PROV_YEAR
    arrange(PROV_YEAR)

  return(df_aggregated)
}

fies_2006_aggregated = add_prov_year_identifier(fies_2006_aggregated)
fies_2009_aggregated = add_prov_year_identifier(fies_2009_aggregated)
fies_2012_aggregated = add_prov_year_identifier(fies_2012_aggregated)
fies_2015_aggregated = add_prov_year_identifier(fies_2015_aggregated)
fies_2018_aggregated = add_prov_year_identifier(fies_2018_aggregated)
```

Add the HHI data from two years before to the FIES data (e.g. the HHI 2004 data is added to the FIES 2006)

```
fies_hhi_2006_aggregated = fies_2006_aggregated %>% mutate(HHI = hhi_map$X2004)
fies_hhi_2009_aggregated = fies_2009_aggregated %>% mutate(HHI = hhi_map$X2007)
fies_hhi_2012_aggregated = fies_2012_aggregated %>% mutate(HHI = hhi_map$X2010)
fies_hhi_2015_aggregated = fies_2015_aggregated %>% mutate(HHI = hhi_map$X2013)
fies_hhi_2018_aggregated = fies_2018_aggregated %>% mutate(HHI = hhi_map$X2016)
```

Row-bind all five datasets

```
fies_hhi_all_years = rbind(fies_hhi_2006_aggregated, fies_hhi_2009_aggregated, fies_hhi_2012_aggregated
                           fies_hhi_2015_aggregated, fies_hhi_2018_aggregated)

View(fies_hhi_all_years)
```

Export to CSV

```
write.csv(fies_hhi_all_years, "fies_hhi_all_years.csv", row.names = FALSE)
```

OLS

Lion De Leon

2024-12-07

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
# install.packages("e1071")
```

```
# Install the necessary libraries
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```
library (corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.92 loaded
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.3.3
```

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3
```

```
library(nortest)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```r
# Store the csv into data
data <- read.csv("fies_hhi_all_years.csv")
```

```r
head(data)
```

```
##                PROV_YEAR    FSIZE    WAGES  NETSHARE CASH_ABROAD CASH_DOMESTIC
## 1              ABRA_2006 4.739583 39148.81 2862.4167   16468.333      4759.267
## 2 AGUSAN DEL NORTE_2006 5.086519 52805.87  729.5151   12976.056      4230.559
## 3   AGUSAN DEL SUR_2006 4.988281 48080.77 1311.4453    2040.344      2254.674
## 4             AKLAN_2006 4.684211 43832.52 1404.6211   13209.621      6957.368
## 5             ALBAY_2006 5.097561 57759.86 1319.2398   15890.488     10238.366
## 6           ANTIQUE_2006 4.625000 31864.91 1713.4375    9686.761      8452.841
##   RENTALS_REC    INTEREST   PENSION DIVIDENDS OTHER_SOURCE NET_RECEIPT    REGFT
## 1   367.08333   0.8333333  3220.504   0.00000      0.00000    4224.196 3928.546
## 2   667.70624 428.2535211  6993.537  26.21127     37.74245    2006.569 4105.980
## 3   472.23958 205.2317708  8033.109  27.34375     13.02083    2928.633 2703.732
## 4    67.36842 209.6105263  1605.695   0.00000      0.00000    3445.616 7019.626
## 5  1441.05691 148.9837398 12072.685  25.35772     16.46341    2844.285 7710.283
## 6   487.15909   8.5227273  1837.500   0.00000      0.00000    3683.188 5968.006
##     NET_CFG  NET_LPR  NET_FISH    NET_FOR   NET_RET   NET_MFG   NET_COM
## 1 13484.504 3803.088  140.0000  145.58333  3570.412  882.2667  910.2250
## 2  7955.911 1436.857  871.8632  467.74849  8627.853 4213.3159  917.7344
## 3 13521.841 2703.924  147.9427 2368.72135  8861.727 1084.2292 1722.0312
## 4  5498.689 1138.763 3240.2684  126.63158  6836.805 1206.4211 1068.3684
## 5  6528.596 1217.577 1453.2866 1122.01016 20553.821 2610.9146 2330.4309
## 6  8709.290 1376.875 2024.5739   61.47727  4703.057  868.2159 2641.4091
##   NET_TRANS    NET_MIN  NET_CONS    NET_NEC    EAINC    TOINC    BREAD     MEAT
## 1  1785.096   20.83333 766.87500    0.00000 25508.88 111983.9 16336.06 8584.138
## 2  3545.243  511.27968  28.16499  385.21730 28912.16 124028.4 16920.10 6834.628
## 3  2682.570 2592.17188   0.00000   72.64323 35745.12 109427.1 16646.52 5888.378
## 4  3360.574    0.00000   0.00000  636.00000 23095.13 109020.8 11944.17 5366.668
## 5  2731.230  313.19919   0.00000 1070.14228 39882.95 166304.2 16285.34 6861.472
## 6  1793.261   68.18182   0.00000  280.26705 22526.61  94232.8 15674.88 3779.631
##       FISH     MILK FOOD_NEC   COFFEE  MINERAL   ALCOHOL   TOBACCO FOOD_HOME
## 1 6115.442 3044.358 5675.421 3026.113 1086.958 1675.8208  778.5250 1298.9375
## 2 6920.781 4399.147 4646.209 4187.489 1283.712 2122.6942  876.1891  897.5533
## 3 6584.794 3264.766 4852.753 3347.549 1232.279 1057.4870  752.5052 1107.3854
## 4 6620.916 2616.984 3489.979 3521.968 1443.153  764.5368  734.0526 1270.4211
## 5 7167.427 4510.539 6104.602 5864.921 1599.852 1304.5224  790.8821 1383.2703
## 6 5329.057 3281.830 4953.807 4208.551 1342.869 1093.5966 1221.4602 1198.9261
##   FOOD_OUTSIDE     FOOD    CLOTH FURNISHING   HEALTH HOUSING_WATER RECREATION
## 1     46045.96 1491.825 47537.79   2507.600 189.9125      1481.008   8598.175
## 2     47974.76 3338.507 51313.27   2934.461 282.7686      3368.006   6814.310
## 3     43679.89 1380.609 45060.49   2114.031 281.1589      2347.424   5533.094
## 4     36157.20 3346.421 39503.62   3019.000 278.1053      4445.484   6250.453
## 5     50405.74 6109.079 56514.82   3192.008 403.0935      6214.516   8250.195
## 6     40271.26 2215.562 42486.82   2204.023 197.0341      2001.585   4699.602
##   EDUCATION OTHER_EXPENDITURE OTHER_DISBURSEMENT    NFOOD     TOTEX    TOTDIS
## 1  5498.975          211.6542           5255.708 3789.154  8050.583 51479.50
## 2  8841.767          543.5875           5237.577 4375.378 12530.726 62491.20
## 3  4690.219          274.4922           3825.188 4268.346 13802.539 43807.73
## 4  6370.389          603.3316           5279.474 3708.468 13736.326 55620.45
## 5  8989.220          584.2256           5775.803 5351.201 19728.774 82272.81
```

```
## 6  4303.091          233.2614          3197.278 3216.051  3947.040 40113.29
##       OTHREC     TOREC FRUITVEG    TRCOM      HHI
## 1  99017.29 107067.87 1375.433 123690.9 58.41424
## 2 113804.46 126335.19 8141.694 135683.9 89.74913
## 3  88868.22 102670.76 5200.435 118760.8 80.73818
## 4  95124.07 108860.39 3681.700 117808.5 76.74162
## 5 138787.63 158516.40 7729.041 184340.1 63.61229
## 6  82600.11  86547.15 2287.307 103553.7 66.31050
```

```r
summary(data)
```

```
##   PROV_YEAR             FSIZE          WAGES           NETSHARE
##  Length:410         Min.   :3.136   Min.   :  8797   Min.   :   0
##  Class :character   1st Qu.:4.478   1st Qu.: 52351   1st Qu.: 887
##  Mode  :character   Median :4.666   Median : 75101   Median :1604
##                     Mean   :4.703   Mean   : 87633   Mean   :1759
##                     3rd Qu.:4.876   3rd Qu.:107590   3rd Qu.:2458
##                     Max.   :6.645   Max.   :309568   Max.   :6988
##   CASH_ABROAD      CASH_DOMESTIC     RENTALS_REC         INTEREST
##  Min.   :  369.9   Min.   :  278.4   Min.   :    0.0   Min.   :   0.00
##  1st Qu.: 9223.4   1st Qu.: 6278.1   1st Qu.:  363.7   1st Qu.:  33.74
##  Median :17389.3   Median : 9797.9   Median :  783.4   Median : 120.86
##  Mean   :20613.3   Mean   :10658.6   Mean   : 1257.4   Mean   : 212.85
##  3rd Qu.:30246.8   3rd Qu.:13537.5   3rd Qu.: 1656.6   3rd Qu.: 262.62
##  Max.   :80107.3   Max.   :37782.6   Max.   :10031.9   Max.   :3580.51
##     PENSION        DIVIDENDS          OTHER_SOURCE      NET_RECEIPT
##  Min.   :    0   Min.   :    0.00   Min.   :   0.00   Min.   :    0
##  1st Qu.: 3507   1st Qu.:    0.62   1st Qu.:   0.00   1st Qu.: 2269
##  Median : 6050   Median :   22.68   Median :  26.17   Median : 3279
##  Mean   : 6605   Mean   :  783.73   Mean   :  89.26   Mean   : 3530
##  3rd Qu.: 9044   3rd Qu.:  123.84   3rd Qu.:  84.92   3rd Qu.: 4447
##  Max.   :31375   Max.   :58109.95   Max.   :1365.14   Max.   :14972
##      REGFT            NET_CFG          NET_LPR              NET_FISH
##  Min.   :  250.6   Min.   :    0   Min.   :   -4.839   Min.   :   -5.03
##  1st Qu.: 4076.4   1st Qu.: 7127   1st Qu.:  987.177   1st Qu.:  666.31
##  Median : 5458.7   Median :11858   Median : 1554.378   Median : 2188.12
##  Mean   : 5888.9   Mean   :14800   Mean   : 2001.275   Mean   : 4098.96
##  3rd Qu.: 6948.4   3rd Qu.:19588   3rd Qu.: 2511.166   3rd Qu.: 4664.53
##  Max.   :24495.8   Max.   :58536   Max.   :15462.000   Max.   :87752.08
##     NET_FOR            NET_RET          NET_MFG           NET_COM
##  Min.   :  -0.878   Min.   :    0   Min.   :-539.3   Min.   :    0.0
##  1st Qu.: 105.282   1st Qu.: 9467   1st Qu.:1121.6   1st Qu.:  211.6
##  Median : 251.980   Median :14987   Median :1753.1   Median : 1690.5
##  Mean   : 425.845   Mean   :15907   Mean   :1987.2   Mean   : 2530.4
##  3rd Qu.: 557.838   3rd Qu.:20012   3rd Qu.:2515.3   3rd Qu.: 3414.3
##  Max.   :3825.649   Max.   :75578   Max.   :9536.8   Max.   :21884.8
##    NET_TRANS        NET_MIN             NET_CONS            NET_NEC
##  Min.   :    0   Min.   :    0.00   Min.   :    0.00   Min.   :    0
##  1st Qu.: 2887   1st Qu.:    0.00   1st Qu.:    0.00   1st Qu.:  254
##  Median : 4616   Median :    0.00   Median :   25.79   Median :  937
##  Mean   : 5378   Mean   :  450.19   Mean   :  477.77   Mean   : 2561
##  3rd Qu.: 7136   3rd Qu.:   91.93   3rd Qu.:  331.67   3rd Qu.: 2895
##  Max.   :21134   Max.   :58650.00   Max.   :34158.65   Max.   :39375
##      EAINC            TOINC            BREAD            MEAT
```

```
##  Min.   : 17851   Min.   : 76691   Min.   :11944   Min.   :  597.1
##  1st Qu.: 38444   1st Qu.:143856   1st Qu.:19333   1st Qu.: 5896.1
##  Median : 46916   Median :186977   Median :22652   Median : 8403.9
##  Mean   : 50617   Mean   :205749   Mean   :22394   Mean   : 9279.8
##  3rd Qu.: 58430   3rd Qu.:249317   3rd Qu.:25443   3rd Qu.:11940.4
##  Max.   :128648   Max.   :504953   Max.   :39368   Max.   :23764.2
##       FISH             MILK             FOOD_NEC          COFFEE
##  Min.   : 3658   Min.   :  553.3   Min.   :  947.3   Min.   : 1038
##  1st Qu.: 7656   1st Qu.: 3161.2   1st Qu.: 1744.0   1st Qu.: 2412
##  Median : 9344   Median : 4271.5   Median : 2582.0   Median : 3184
##  Mean   : 9750   Mean   : 4578.4   Mean   : 3540.5   Mean   : 3700
##  3rd Qu.:11418   3rd Qu.: 5718.0   3rd Qu.: 5029.8   3rd Qu.: 4434
##  Max.   :23799   Max.   :20874.7   Max.   :14139.2   Max.   :13141
##     MINERAL          ALCOHOL          TOBACCO           FOOD_HOME
##  Min.   : 449.1   Min.   :   0.0   Min.   :   9.264   Min.   :    574.6
##  1st Qu.:1412.4   1st Qu.: 950.2   1st Qu.:1004.590   1st Qu.:   1387.5
##  Median :1758.1   Median :1242.6   Median :1401.925   Median :  58443.9
##  Mean   :1982.1   Mean   :1392.2   Mean   :1635.036   Mean   :  42319.1
##  3rd Qu.:2306.3   3rd Qu.:1662.9   3rd Qu.:2018.066   3rd Qu.:  70269.6
##  Max.   :5480.2   Max.   :4900.8   Max.   :6812.276   Max.   :113198.1
##   FOOD_OUTSIDE            FOOD               CLOTH          FURNISHING
##  Min.   :    36.36   Min.   :    808.5   Min.   :   1481   Min.   :  560.3
##  1st Qu.:  7459.55   1st Qu.:   5127.5   1st Qu.:   4349   1st Qu.: 2721.6
##  Median : 18161.34   Median :  64967.7   Median :   6464   Median : 3633.2
##  Mean   : 28491.91   Mean   :  51188.1   Mean   :  26875   Mean   : 3937.2
##  3rd Qu.: 48151.84   3rd Qu.:  79115.0   3rd Qu.:  51298   3rd Qu.: 4754.7
##  Max.   :103282.61   Max.   :154160.8   Max.   :121330   Max.   :12562.0
##      HEALTH          HOUSING_WATER        RECREATION      EDUCATION
##  Min.   :   98.09   Min.   :    473.2   Min.   :  292   Min.   :    948.5
##  1st Qu.:  282.95   1st Qu.:   4396.8   1st Qu.: 1449   1st Qu.:   4849.7
##  Median : 3576.87   Median :  19607.5   Median : 2502   Median :   6410.8
##  Mean   : 3869.69   Mean   :  21653.5   Mean   : 4756   Mean   :   7900.5
##  3rd Qu.: 6566.19   3rd Qu.:  31172.1   3rd Qu.: 7060   3rd Qu.:   9069.2
##  Max.   :25943.30   Max.   :115487.4   Max.   :27121   Max.   :100566.6
##  OTHER_EXPENDITURE OTHER_DISBURSEMENT      NFOOD              TOTEX
##  Min.   :   79.5   Min.   :  111.8    Min.   :    914.6   Min.   :      0
##  1st Qu.:  611.1   1st Qu.: 5885.9    1st Qu.:   4938.1   1st Qu.: 19068
##  Median : 2395.1   Median :13310.4    Median : 66315.4   Median :131223
##  Mean   : 3193.3   Mean   :17304.6    Mean   : 62795.7   Mean   :116777
##  3rd Qu.: 4184.7   3rd Qu.:25399.7    3rd Qu.: 98170.0   3rd Qu.:176482
##  Max.   :24222.7   Max.   :87109.4    Max.   :265146.4   Max.   :403508
##      TOTDIS           OTHREC              TOREC            FRUITVEG
##  Min.   : 28541   Min.   :   186.2   Min.   : 68866   Min.   : 1375
##  1st Qu.: 76386   1st Qu.: 13414.7   1st Qu.:142330   1st Qu.: 6430
##  Median :152093   Median : 24944.9   Median :199785   Median : 7904
##  Mean   :154960   Mean   : 64602.3   Mean   :213060   Mean   : 9166
##  3rd Qu.:203396   3rd Qu.:111219.7   3rd Qu.:264317   3rd Qu.:10423
##  Max.   :476092   Max.   :348933.4   Max.   :551718   Max.   :61160
##      TRCOM              HHI
##  Min.   :  4401   Min.   : 25.11
##  1st Qu.: 12060   1st Qu.: 57.05
##  Median : 20965   Median : 75.32
##  Mean   : 79330   Mean   : 90.33
##  3rd Qu.:145831   3rd Qu.:107.64
```

```
##  Max.   :455336   Max.   :320.25
```

```r
# Removing non-numerical columns
reg_data <- data[-1]
```

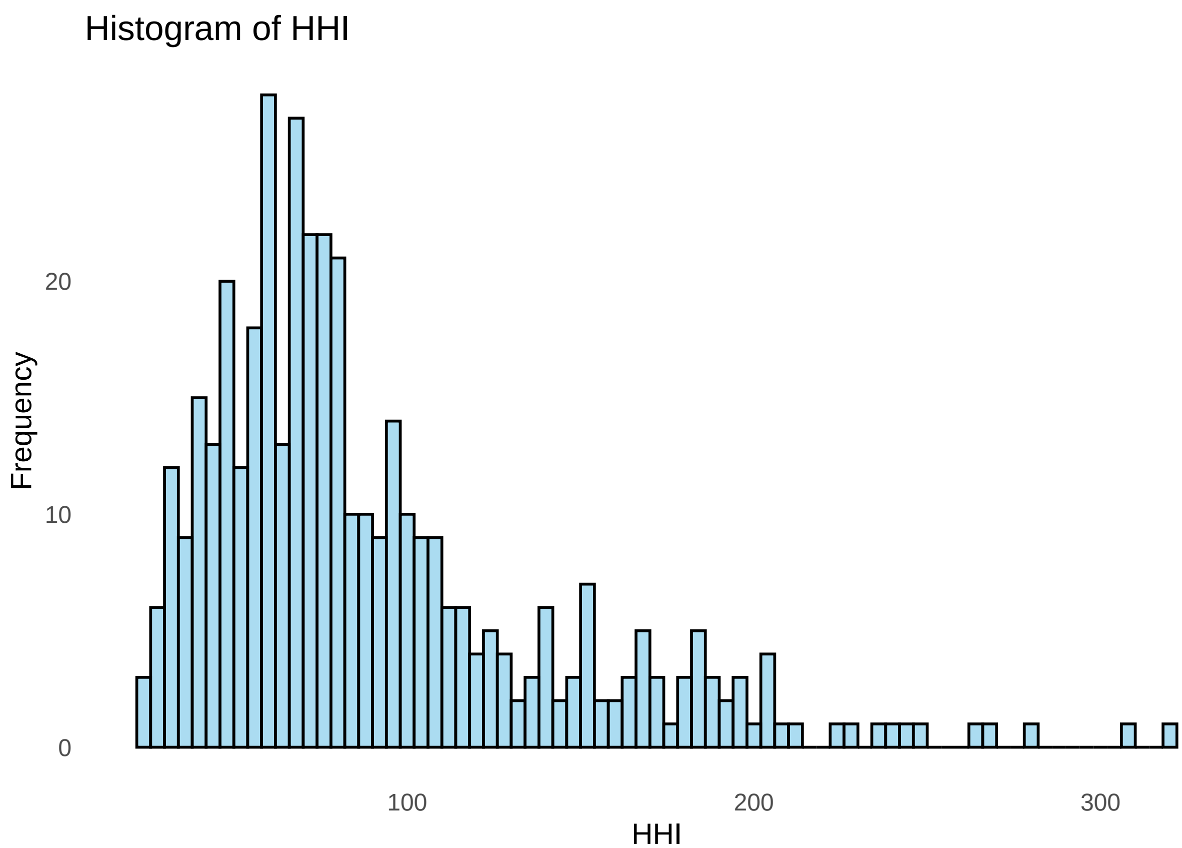
```r
class(data)
```

```
## [1] "data.frame"
```

```r
# Calculate skewness for each numeric column in the dataset
reg_skew_values <- sapply(reg_data, function(x) if (is.numeric(x)) skewness(x, na.rm = TRUE) else NA)
reg_skewed_vars <- reg_skew_values[abs(reg_skew_values) > 1]
# Skewness 1 is commonly chosen in practice
names(reg_skewed_vars)
```

```
##  [1] "WAGES"              "NETSHARE"            "RENTALS_REC"
##  [4] "INTEREST"           "PENSION"             "DIVIDENDS"
##  [7] "OTHER_SOURCE"       "NET_RECEIPT"         "REGFT"
## [10] "NET_CFG"            "NET_LPR"             "NET_FISH"
## [13] "NET_FOR"            "NET_RET"             "NET_MFG"
## [16] "NET_COM"            "NET_TRANS"           "NET_MIN"
## [19] "NET_CONS"           "NET_NEC"             "EAINC"
## [22] "TOINC"              "FISH"                "MILK"
## [25] "FOOD_NEC"           "COFFEE"              "MINERAL"
## [28] "ALCOHOL"            "TOBACCO"             "FURNISHING"
## [31] "HEALTH"             "HOUSING_WATER"       "RECREATION"
## [34] "EDUCATION"          "OTHER_EXPENDITURE"   "OTHER_DISBURSEMENT"
## [37] "OTHREC"             "FRUITVEG"            "TRCOM"
## [40] "HHI"
```

```r
# Visualizing the skewness of HHI
skewness_value <- skewness(data$HHI, na.rm = TRUE)  # na.rm = TRUE ignores NA values
ggplot(data, aes(x = HHI)) +
  geom_histogram(binwidth = 4, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histogram of HHI", x = "HHI", y = "Frequency") +
  theme_minimal() +
  theme(panel.grid = element_blank())
```




```r
# Replacing the skewed variables with their log transformation
for (col_name in names(reg_skewed_vars)) {
  reg_data[[paste("LOG", col_name, sep = "_")]] <- log1p(reg_data[[col_name]])
  reg_data[[col_name]] <- NULL
}
```

```
## Warning in log1p(reg_data[[col_name]]): NaNs produced

## Warning in log1p(reg_data[[col_name]]): NaNs produced

## Warning in log1p(reg_data[[col_name]]): NaNs produced
```

```r
head(reg_data)
```

```
##      FSIZE CASH_ABROAD CASH_DOMESTIC    BREAD     MEAT FOOD_HOME FOOD_OUTSIDE
## 1 4.739583   16468.333      4759.267 16336.06 8584.138 1298.9375     46045.96
## 2 5.086519   12976.056      4230.559 16920.10 6834.628  897.5533     47974.76
## 3 4.988281    2040.344      2254.674 16646.52 5888.378 1107.3854     43679.89
## 4 4.684211   13209.621      6957.368 11944.17 5366.668 1270.4211     36157.20
## 5 5.097561   15890.488     10238.366 16285.34 6861.472 1383.2703     50405.74
## 6 4.625000    9686.761      8452.841 15674.88 3779.631 1198.9261     40271.26
##       FOOD    CLOTH    NFOOD     TOTEX   TOTDIS     TOREC LOG_WAGES
## 1 1491.825 47537.79 3789.154  8050.583 51479.50 107067.87  10.57515
## 2 3338.507 51313.27 4375.378 12530.726 62491.20 126335.19  10.87440
## 3 1380.609 45060.49 4268.346 13802.539 43807.73 102670.76  10.78066
```

```
## 4 3346.421 39503.62 3708.468 13736.326 55620.45 108860.39  10.68815
## 5 6109.079 56514.82 5351.201 19728.774 82272.81 158516.40  10.96407
## 6 2215.562 42486.82 3216.051  3947.040 40113.29  86547.15  10.36929
##   LOG_NETSHARE LOG_RENTALS_REC LOG_INTEREST LOG_PENSION LOG_DIVIDENDS
## 1     7.959771        5.908309    0.6061358    8.077604      0.000000
## 2     6.593750        6.505345    6.0620477    8.852885      3.303631
## 3     7.179647        6.159602    5.3290006    8.991451      3.344407
## 4     7.248235        4.224911    5.3500106    7.381934      0.000000
## 5     7.185569        7.273826    5.0105269    9.398784      3.271761
## 6     7.446840        6.190641    2.2536813    7.516705      0.000000
##   LOG_OTHER_SOURCE LOG_NET_RECEIPT LOG_REGFT LOG_NET_CFG LOG_NET_LPR
## 1         0.000000        8.348821  8.276279    9.509371    8.243831
## 2         3.656936        7.604680  8.320443    8.981796    7.270909
## 3         2.640544        7.982632  7.902758    9.512135    7.902829
## 4         0.000000        8.145148  8.856608    8.612447    7.038576
## 5         2.860108        7.953418  8.950440    8.784100    7.105439
## 6         0.000000        8.211805  8.694336    9.072260    7.228298
##   LOG_NET_FISH LOG_NET_FOR LOG_NET_RET LOG_NET_MFG LOG_NET_COM LOG_NET_TRANS
## 1     4.948760    4.987594    8.180716    6.783627    6.814790      7.487787
## 2     6.771779    6.150066    9.062867    8.346243    6.822997      8.173644
## 3     5.003562    7.770528    9.089610    6.989546    7.451840      7.894903
## 4     8.083720    4.849148    8.830222    7.096242    6.974823      8.120165
## 5     7.282271    7.023768    9.930851    7.867839    7.754237      7.912873
## 6     7.613608    4.134803    8.456181    6.767592    7.879446      7.492349
##   LOG_NET_MIN LOG_NET_CONS LOG_NET_NEC LOG_EAINC LOG_TOINC LOG_FISH LOG_MILK
## 1    3.083438     6.643627    0.000000  10.14682  11.62612 8.718736 8.021374
## 2    6.238871     3.372969    5.956400  10.27205  11.72827 8.842428 8.389393
## 3    7.860637     0.000000    4.299232  10.48420  11.60302 8.792670 8.091250
## 4    0.000000     0.000000    6.456770  10.04742  11.59930 8.798140 7.870160
## 5    5.750027     0.000000    6.976481  10.59373  12.02158 8.877441 8.414394
## 6    4.236738     0.000000    5.639305  10.02250  11.45353 8.581117 8.096461
##   LOG_FOOD_NEC LOG_COFFEE LOG_MINERAL LOG_ALCOHOL LOG_TOBACCO LOG_FURNISHING
## 1     8.644076   8.015364    6.992058    7.424655    6.658685       7.827480
## 2     8.444022   8.340095    7.158290    7.660912    6.776723       7.984620
## 3     8.487507   8.116283    7.117431    6.964596    6.624736       7.656825
## 4     8.157937   8.167059    7.275278    6.640577    6.599942       8.013012
## 5     8.716962   8.676915    7.378291    7.174359    6.674413       8.068719
## 6     8.508113   8.345111    7.203308    6.998141    7.108621       7.698493
##   LOG_HEALTH LOG_HOUSING_WATER LOG_RECREATION LOG_EDUCATION
## 1   5.251815          7.301153       9.059422      8.612499
## 2   5.648159          8.122373       8.826927      9.087355
## 3   5.642470          7.761500       8.618683      8.453448
## 4   5.631589          8.399869       8.740569      8.759573
## 5   6.001646          8.734804       9.018113      9.103893
## 6   5.288439          7.602194       8.455446      8.367321
##   LOG_OTHER_EXPENDITURE LOG_OTHER_DISBURSEMENT LOG_OTHREC LOG_FRUITVEG
## 1              5.359667               8.567260   11.50306     7.227251
## 2              6.300029               8.563805   11.64225     9.004876
## 3              5.618559               8.249624   11.39492     8.556690
## 4              6.404123               8.571771   11.46295     8.211401
## 5              6.371997               8.661606   11.84071     8.952869
## 6              5.456437               8.070368   11.32178     7.735567
##   LOG_TRCOM  LOG_HHI
## 1  11.72555 4.084534
```

```
## 2  11.81809 4.508099
## 3  11.68487 4.403521
## 4  11.67682 4.353391
## 5  12.12454 4.168405
## 6  11.54786 4.209316
```

```r
names(reg_data)
```

```
##  [1] "FSIZE"                  "CASH_ABROAD"            "CASH_DOMESTIC"
##  [4] "BREAD"                  "MEAT"                   "FOOD_HOME"
##  [7] "FOOD_OUTSIDE"           "FOOD"                   "CLOTH"
## [10] "NFOOD"                  "TOTEX"                  "TOTDIS"
## [13] "TOREC"                  "LOG_WAGES"              "LOG_NETSHARE"
## [16] "LOG_RENTALS_REC"        "LOG_INTEREST"           "LOG_PENSION"
## [19] "LOG_DIVIDENDS"          "LOG_OTHER_SOURCE"       "LOG_NET_RECEIPT"
## [22] "LOG_REGFT"              "LOG_NET_CFG"            "LOG_NET_LPR"
## [25] "LOG_NET_FISH"           "LOG_NET_FOR"            "LOG_NET_RET"
## [28] "LOG_NET_MFG"            "LOG_NET_COM"            "LOG_NET_TRANS"
## [31] "LOG_NET_MIN"            "LOG_NET_CONS"           "LOG_NET_NEC"
## [34] "LOG_EAINC"              "LOG_TOINC"              "LOG_FISH"
## [37] "LOG_MILK"               "LOG_FOOD_NEC"           "LOG_COFFEE"
## [40] "LOG_MINERAL"            "LOG_ALCOHOL"            "LOG_TOBACCO"
## [43] "LOG_FURNISHING"         "LOG_HEALTH"             "LOG_HOUSING_WATER"
## [46] "LOG_RECREATION"         "LOG_EDUCATION"          "LOG_OTHER_EXPENDITURE"
## [49] "LOG_OTHER_DISBURSEMENT" "LOG_OTHREC"             "LOG_FRUITVEG"
## [52] "LOG_TRCOM"              "LOG_HHI"
```

```r
# Regression for the VIF
reg_data.done = lm(LOG_HHI~., data = reg_data)

# ---------------------------------
# From here on, all comments at the bottom are for skewness threshold = 2
#reg_data.done = lm(HHI~., data = reg_data)
```

```r
# Checking multicollinearity using VIF
vif(reg_data.done)
```

```
##                  FSIZE            CASH_ABROAD          CASH_DOMESTIC
##               3.079540               6.316105               4.271934
##                  BREAD                   MEAT              FOOD_HOME
##              16.912521              29.855397            1833.393807
##           FOOD_OUTSIDE                   FOOD                  CLOTH
##             210.082396             761.436440             184.976930
##                  NFOOD                  TOTEX                 TOTDIS
##             246.742356             611.654765             205.889406
##                  TOREC              LOG_WAGES           LOG_NETSHARE
##              94.717243              28.008132               2.633396
##        LOG_RENTALS_REC           LOG_INTEREST            LOG_PENSION
##               2.380214               1.513195               3.061368
##          LOG_DIVIDENDS       LOG_OTHER_SOURCE        LOG_NET_RECEIPT
##               1.650626               1.468229               9.161819
##              LOG_REGFT            LOG_NET_CFG            LOG_NET_LPR
```

```
##              4.283948              5.183623              4.074911
##          LOG_NET_FISH           LOG_NET_FOR           LOG_NET_RET
##              2.084246              2.335774              2.913310
##           LOG_NET_MFG           LOG_NET_COM         LOG_NET_TRANS
##              2.449573              4.199243              2.733062
##           LOG_NET_MIN          LOG_NET_CONS           LOG_NET_NEC
##              1.594583              1.775183              2.449931
##             LOG_EAINC             LOG_TOINC              LOG_FISH
##              8.603564             85.506385              8.545695
##              LOG_MILK          LOG_FOOD_NEC            LOG_COFFEE
##              9.002129             24.620761             10.573167
##           LOG_MINERAL           LOG_ALCOHOL           LOG_TOBACCO
##              5.755506              5.059240              4.040691
##         LOG_FURNISHING            LOG_HEALTH     LOG_HOUSING_WATER
##              8.153985             27.822106             32.566812
##         LOG_RECREATION         LOG_EDUCATION LOG_OTHER_EXPENDITURE
##             19.804215              6.274300             11.258865
## LOG_OTHER_DISBURSEMENT            LOG_OTHREC          LOG_FRUITVEG
##             11.920122             19.222580              3.989574
##             LOG_TRCOM
##             71.688707
```

```r
# Selecting variables with VIF < 5
reg_data.done = lm(LOG_HHI~FSIZE + CASH_DOMESTIC + LOG_NETSHARE + LOG_RENTALS_REC + LOG_INTEREST + LOG_I
summary(reg_data.done)
```

```
##
## Call:
## lm(formula = LOG_HHI ~ FSIZE + CASH_DOMESTIC + LOG_NETSHARE +
##     LOG_RENTALS_REC + LOG_INTEREST + LOG_PENSION + LOG_DIVIDENDS +
##     LOG_OTHER_SOURCE + LOG_REGFT + LOG_NET_LPR + LOG_NET_FISH +
##     LOG_NET_FOR + LOG_NET_RET + LOG_NET_MFG + LOG_NET_COM + LOG_NET_TRANS +
##     LOG_NET_MIN + LOG_NET_CONS + LOG_NET_NEC + LOG_TOBACCO +
##     LOG_FRUITVEG, data = reg_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15246 -0.28662  0.00025  0.28980  1.21805
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.964e+00  8.056e-01   6.161 1.82e-09 ***
## FSIZE            -4.299e-02  7.294e-02  -0.589 0.555888
## CASH_DOMESTIC     3.046e-06  4.908e-06   0.621 0.535167
## LOG_NETSHARE     -5.330e-02  2.322e-02  -2.296 0.022220 *
## LOG_RENTALS_REC  -1.792e-02  1.753e-02  -1.022 0.307358
## LOG_INTEREST     -3.251e-03  1.307e-02  -0.249 0.803752
## LOG_PENSION      -6.085e-02  2.646e-02  -2.299 0.022018 *
## LOG_DIVIDENDS     1.061e-02  9.707e-03   1.094 0.274848
## LOG_OTHER_SOURCE -1.084e-02  1.159e-02  -0.935 0.350323
## LOG_REGFT         2.638e-01  5.595e-02   4.716 3.37e-06 ***
## LOG_NET_LPR      -1.127e-01  2.376e-02  -4.742 2.98e-06 ***
## LOG_NET_FISH     -2.818e-02  1.140e-02  -2.473 0.013824 *
## LOG_NET_FOR      -3.139e-02  1.383e-02  -2.270 0.023745 *
```

```
## LOG_NET_RET      -8.626e-02  3.716e-02  -2.321 0.020783 *
## LOG_NET_MFG       3.930e-02  2.862e-02   1.373 0.170572
## LOG_NET_COM      -3.025e-02  9.121e-03  -3.316 0.000999 ***
## LOG_NET_TRANS    -4.639e-02  3.206e-02  -1.447 0.148719
## LOG_NET_MIN       3.901e-03  8.687e-03   0.449 0.653594
## LOG_NET_CONS     -2.260e-02  8.960e-03  -2.522 0.012069 *
## LOG_NET_NEC      -6.481e-02  1.247e-02  -5.195 3.32e-07 ***
## LOG_TOBACCO       4.479e-02  3.943e-02   1.136 0.256608
## LOG_FRUITVEG      8.450e-02  4.919e-02   1.718 0.086597 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4306 on 385 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.3162, Adjusted R-squared:  0.2789
## F-statistic: 8.477 on 21 and 385 DF,  p-value: < 2.2e-16
```

```r
#reg_data.done = lm(HHI~FSIZE + NETSHARE + PENSION + NET_RECEIPT + NET_MFG + NET_TRANS + ALCOHOL + TOBA
#summary(reg_data.done)
```

```r
# Variable selection by choosing significant variables with alpha = 0.05
reg_data_reduced.done = lm(LOG_HHI~LOG_NETSHARE + LOG_PENSION + LOG_REGFT + LOG_NET_LPR +  LOG_NET_FISH
summary(reg_data_reduced.done)
```

```
##
## Call:
## lm(formula = LOG_HHI ~ LOG_NETSHARE + LOG_PENSION + LOG_REGFT +
##     LOG_NET_LPR + LOG_NET_FISH + LOG_NET_FOR + LOG_NET_RET +
##     LOG_NET_COM + LOG_NET_CONS + LOG_NET_NEC, data = reg_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15403 -0.27421  0.00109  0.29346  1.25761
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.120468   0.432362  11.843  < 2e-16 ***
## LOG_NETSHARE -0.052349   0.021960  -2.384 0.017604 *
## LOG_PENSION  -0.050866   0.022639  -2.247 0.025204 *
## LOG_REGFT     0.316038   0.049394   6.398 4.43e-10 ***
## LOG_NET_LPR  -0.115811   0.022386  -5.173 3.66e-07 ***
## LOG_NET_FISH -0.026556   0.010662  -2.491 0.013157 *
## LOG_NET_FOR  -0.031085   0.012995  -2.392 0.017219 *
## LOG_NET_RET  -0.082294   0.032621  -2.523 0.012036 *
## LOG_NET_COM  -0.030077   0.008003  -3.758 0.000197 ***
## LOG_NET_CONS -0.025688   0.008506  -3.020 0.002691 **
## LOG_NET_NEC  -0.067493   0.010764  -6.271 9.40e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4316 on 397 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.2919, Adjusted R-squared:  0.2741
## F-statistic: 16.37 on 10 and 397 DF,  p-value: < 2.2e-16
```

```
#reg_data_reduced.done = lm(HHI~NET_RECEIPT + NET_TRANS + ALCOHOL + LOG_REGFT + LOG_NET_LPR + LOG_NET_F
#summary(reg_data_reduced.done)
```

```
# ADF Test for normality assumption
ad.test(reg_data_reduced.done$residuals)
```

```
##
##  Anderson-Darling normality test
##
## data:  reg_data_reduced.done$residuals
## A = 0.18341, p-value = 0.9098
```

```
# Saving the file
# write.csv(reg_data, file = "regression_data.csv", row.names = FALSE)
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

HHI_SpatialReg

2024-12-08

```r
library(fuzzyjoin)
```

```
## Warning: package 'fuzzyjoin' was built under R version 4.4.2
```

```r
library(sf)
```

```
## Warning: package 'sf' was built under R version 4.4.2
```

```
## Linking to GEOS 3.12.2, GDAL 3.9.3, PROJ 9.4.1; sf_use_s2() is TRUE
```

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(stringdist)
#install.packages(c('fuzzyjoin', 'sf'))

#Library for Spatial Regression
library(spdep)
```

```
## Warning: package 'spdep' was built under R version 4.4.2
```

```
## Loading required package: spData
```

```
## Warning: package 'spData' was built under R version 4.4.2
```

```
## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source')`
```

```r
library(spatialreg)
```

```
## Warning: package 'spatialreg' was built under R version 4.4.2

## Loading required package: Matrix

##
## Attaching package: 'spatialreg'

## The following objects are masked from 'package:spdep':
##
##     get.ClusterOption, get.coresOption, get.mcOption,
##     get.VerboseOption, get.ZeroPolicyOption, set.ClusterOption,
##     set.coresOption, set.mcOption, set.VerboseOption,
##     set.ZeroPolicyOption
```

```r
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.4.2
```

```r
#Libraries for Spatial Data
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.2

## Warning: package 'ggplot2' was built under R version 4.4.2

## Warning: package 'readr' was built under R version 4.4.2

## Warning: package 'stringr' was built under R version 4.4.2

## Warning: package 'forcats' was built under R version 4.4.2

## Warning: package 'lubridate' was built under R version 4.4.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x tidyr::expand()  masks Matrix::expand()
## x tidyr::extract() masks stringdist::extract()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x tidyr::pack()    masks Matrix::pack()
## x tidyr::unpack()  masks Matrix::unpack()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(colorspace)
```

```
## Warning: package 'colorspace' was built under R version 4.4.2
```

```r
library(rnaturalearth)
```

```
## Warning: package 'rnaturalearth' was built under R version 4.4.2
```

```r
library(sf)
library(devtools)
```

```
## Warning: package 'devtools' was built under R version 4.4.2

## Loading required package: usethis
```

```r
library(rnaturalearthhires)
library(flextable)
```

```
## Warning: package 'flextable' was built under R version 4.4.2

##
## Attaching package: 'flextable'
##
## The following object is masked from 'package:purrr':
##
##     compose
```

```r
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```r
install.packages("flextable")
```

```
## Warning: package 'flextable' is in use and will not be installed
```

```r
install.packages("stargazer")
```

```
## Warning: package 'stargazer' is in use and will not be installed
```

```r
#library(readstata13)
```

```r
ph <- st_read("hhi_map.shp")
```

```
## Reading layer `hhi_map' from data source
##   `C:\Users\Aldie\Documents\Academics\00 Y4S1 Files\02 ECON 185.78i - EconDSci\Groupworkj\Regression
##   using driver `ESRI Shapefile'
## Simple feature collection with 85 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 114.2779 ymin: 4.587294 xmax: 126.605 ymax: 21.12189
## Geodetic CRS:  WGS 84
```

```r
ph <- st_as_sf(ph)

#Rename the column in shp file
ph <- rename(ph, Province = X)
head(ph)
```

```
## Simple feature collection with 6 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 118.0642 ymin: 4.587294 xmax: 125.9392 ymax: 8.163486
## Geodetic CRS:  WGS 84
##          Province     X2004     X2007     X2010     X2013     X2016      lon
## 1   LANAO DEL SUR  46.84711  49.89349  68.73035  74.60973  67.80187 124.3362
## 2     MAGUINDANAO 126.93136 150.12959 152.62346 115.27281 102.51912 124.3847
## 3            SULU  85.17795  91.14583  87.68201 141.89609  96.14512 121.0543
## 4         BASILAN 137.90100 144.79500  90.66358  97.50297 121.16571 122.0291
## 5       TAWI-TAWI 109.56903 109.09091 133.92857 131.82160 153.89351 119.9010
## 6 DAVAO DEL NORTE  88.68584  88.38384  90.67952  90.67952  91.82736 125.6423
##        lat                       geometry
## 1 7.796685 MULTIPOLYGON (((124.484 8.1...
## 2 7.031549 MULTIPOLYGON (((124.047 7.3...
## 3 5.953217 MULTIPOLYGON (((120.6989 6....
## 4 6.565456 MULTIPOLYGON (((121.5931 6....
## 5 5.239002 MULTIPOLYGON (((118.4204 7....
## 6 7.585208 MULTIPOLYGON (((125.3794 8....
```

```r
colnames(data)
```

```
## NULL
```

```r
data <- read.csv("fies_hhi_all_years.csv")
head(data)
```

```
##                PROV_YEAR    FSIZE    WAGES  NETSHARE CASH_ABROAD CASH_DOMESTIC
## 1              ABRA_2006 4.739583 39148.81 2862.4167   16468.333      4759.267
## 2 AGUSAN DEL NORTE_2006 5.086519 52805.87  729.5151   12976.056      4230.559
## 3   AGUSAN DEL SUR_2006 4.988281 48080.77 1311.4453    2040.344      2254.674
## 4             AKLAN_2006 4.684211 43832.52 1404.6211   13209.621      6957.368
## 5             ALBAY_2006 5.097561 57759.86 1319.2398   15890.488     10238.366
## 6           ANTIQUE_2006 4.625000 31864.91 1713.4375    9686.761      8452.841
```

```
##   RENTALS_REC    INTEREST   PENSION DIVIDENDS OTHER_SOURCE NET_RECEIPT    REGFT
## 1   367.08333   0.8333333  3220.504   0.00000      0.00000    4224.196 3928.546
## 2   667.70624 428.2535211  6993.537  26.21127     37.74245    2006.569 4105.980
## 3   472.23958 205.2317708  8033.109  27.34375     13.02083    2928.633 2703.732
## 4    67.36842 209.6105263  1605.695   0.00000      0.00000    3445.616 7019.626
## 5  1441.05691 148.9837398 12072.685  25.35772     16.46341    2844.285 7710.283
## 6   487.15909   8.5227273  1837.500   0.00000      0.00000    3683.188 5968.006
##     NET_CFG  NET_LPR  NET_FISH    NET_FOR    NET_RET   NET_MFG   NET_COM
## 1 13484.504 3803.088  140.0000  145.58333  3570.412  882.2667  910.2250
## 2  7955.911 1436.857  871.8632  467.74849  8627.853 4213.3159  917.7344
## 3 13521.841 2703.924  147.9427 2368.72135  8861.727 1084.2292 1722.0312
## 4  5498.689 1138.763 3240.2684  126.63158  6836.805 1206.4211 1068.3684
## 5  6528.596 1217.577 1453.2866 1122.01016 20553.821 2610.9146 2330.4309
## 6  8709.290 1376.875 2024.5739   61.47727  4703.057  868.2159 2641.4091
##   NET_TRANS    NET_MIN  NET_CONS    NET_NEC    EAINC    TOINC    BREAD     MEAT
## 1  1785.096   20.83333 766.87500    0.00000 25508.88 111983.9 16336.06 8584.138
## 2  3545.243  511.27968  28.16499  385.21730 28912.16 124028.4 16920.10 6834.628
## 3  2682.570 2592.17188   0.00000   72.64323 35745.12 109427.1 16646.52 5888.378
## 4  3360.574    0.00000   0.00000  636.00000 23095.13 109020.8 11944.17 5366.668
## 5  2731.230  313.19919   0.00000 1070.14228 39882.95 166304.2 16285.34 6861.472
## 6  1793.261   68.18182   0.00000  280.26705 22526.61  94232.8 15674.88 3779.631
##       FISH     MILK FOOD_NEC   COFFEE  MINERAL   ALCOHOL   TOBACCO FOOD_HOME
## 1 6115.442 3044.358 5675.421 3026.113 1086.958 1675.8208  778.5250 1298.9375
## 2 6920.781 4399.147 4646.209 4187.489 1283.712 2122.6942  876.1891  897.5533
## 3 6584.794 3264.766 4852.753 3347.549 1232.279 1057.4870  752.5052 1107.3854
## 4 6620.916 2616.984 3489.979 3521.968 1443.153  764.5368  734.0526 1270.4211
## 5 7167.427 4510.539 6104.602 5864.921 1599.852 1304.5224  790.8821 1383.2703
## 6 5329.057 3281.830 4953.807 4208.551 1342.869 1093.5966 1221.4602 1198.9261
##   FOOD_OUTSIDE     FOOD    CLOTH FURNISHING   HEALTH HOUSING_WATER RECREATION
## 1     46045.96 1491.825 47537.79   2507.600 189.9125      1481.008   8598.175
## 2     47974.76 3338.507 51313.27   2934.461 282.7686      3368.006   6814.310
## 3     43679.89 1380.609 45060.49   2114.031 281.1589      2347.424   5533.094
## 4     36157.20 3346.421 39503.62   3019.000 278.1053      4445.484   6250.453
## 5     50405.74 6109.079 56514.82   3192.008 403.0935      6214.516   8250.195
## 6     40271.26 2215.562 42486.82   2204.023 197.0341      2001.585   4699.602
##   EDUCATION OTHER_EXPENDITURE OTHER_DISBURSEMENT    NFOOD     TOTEX   TOTDIS
## 1  5498.975          211.6542           5255.708 3789.154  8050.583 51479.50
## 2  8841.767          543.5875           5237.577 4375.378 12530.726 62491.20
## 3  4690.219          274.4922           3825.188 4268.346 13802.539 43807.73
## 4  6370.389          603.3316           5279.474 3708.468 13736.326 55620.45
## 5  8989.220          584.2256           5775.803 5351.201 19728.774 82272.81
## 6  4303.091          233.2614           3197.278 3216.051  3947.040 40113.29
##      OTHREC     TOREC FRUITVEG    TRCOM      HHI
## 1  99017.29 107067.87 1375.433 123690.9 58.41424
## 2 113804.46 126335.19 8141.694 135683.9 89.74913
## 3  88868.22 102670.76 5200.435 118760.8 80.73818
## 4  95124.07 108860.39 3681.700 117808.5 76.74162
## 5 138787.63 158516.40 7729.041 184340.1 63.61229
## 6  82600.11  86547.15 2287.307 103553.7 66.31050
```

```r
#Consider the variables which are significant from OLS regression
fies_subs <- data %>%
  select(PROV_YEAR, NETSHARE, PENSION, REGFT, NET_LPR, NET_FISH, NET_FOR, NET_RET, NET_COM, NET_CONS, N
head(fies_subs, 5)
```

```
##                PROV_YEAR  NETSHARE   PENSION    REGFT  NET_LPR  NET_FISH
## 1              ABRA_2006 2862.4167  3220.504 3928.546 3803.088  140.0000
## 2 AGUSAN DEL NORTE_2006  729.5151  6993.537 4105.980 1436.857  871.8632
## 3   AGUSAN DEL SUR_2006 1311.4453  8033.109 2703.732 2703.924  147.9427
## 4             AKLAN_2006 1404.6211  1605.695 7019.626 1138.763 3240.2684
## 5             ALBAY_2006 1319.2398 12072.685 7710.283 1217.577 1453.2866
##     NET_FOR   NET_RET   NET_COM  NET_CONS    NET_NEC      HHI
## 1  145.5833  3570.412  910.2250 766.87500    0.00000 58.41424
## 2  467.7485  8627.853  917.7344  28.16499  385.21730 89.74913
## 3 2368.7214  8861.727 1722.0312   0.00000   72.64323 80.73818
## 4  126.6316  6836.805 1068.3684   0.00000  636.00000 76.74162
## 5 1122.0102 20553.821 2330.4309   0.00000 1070.14228 63.61229
```

```
#Do log
fies_subs <- fies_subs %>%
  mutate_at(vars(NETSHARE, PENSION, REGFT, NET_LPR, NET_FISH, NET_FOR, NET_RET, NET_COM, NET_CONS, NET_
```

```
## Warning: `funs()` was deprecated in dplyr 0.8.0.
## i Please use a list of either functions or lambdas:
##
## # Simple named list: list(mean = mean, median = median)
##
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)
##
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
## Warning: There were 2 warnings in `mutate()`.
## The first warning was:
## i In argument: `NET_LPR = log(NET_LPR + 1)`.
## Caused by warning in `log()`:
## ! NaNs produced
## i Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

```
fies_subs[is.na(fies_subs)] <- 0

#Extract province name from the subset
fies_subs$Province <- sapply(strsplit(fies_subs$PROV_YEAR, "_"), `[`, 1)
```

```
fies_aggregated <- fies_subs %>%
  group_by(Province) %>%
  summarize(
    AVG_NETSHARE = mean(NETSHARE),
    AVG_PENSION = mean(PENSION),
    AVG_REGFT = mean(REGFT),
    AVG_NET_LPR = mean(NET_LPR),
    AVG_NET_FISH = mean(NET_FISH),
    AVG_NET_FOR = mean(NET_FOR),
    AVG_NET_RET = mean(NET_RET),
    AVG_NET_COM = mean(NET_COM),
    AVG_NET_CONS = mean(NET_CONS),
```

```r
    AVG_NET_NEC = mean(NET_NEC),
    AVG_HHI = mean(HHI)
  )
head(fies_aggregated, 8)
```

```
## # A tibble: 8 x 12
##   Province          AVG_NETSHARE AVG_PENSION AVG_REGFT AVG_NET_LPR AVG_NET_FISH
##   <chr>                    <dbl>       <dbl>     <dbl>       <dbl>        <dbl>
## 1 ABRA                      7.95        8.66      8.37        8.03         5.95
## 2 AGUSAN DEL NORTE          7.13        9.22      8.39        7.28         6.62
## 3 AGUSAN DEL SUR            7.05        8.73      8.10        7.30         4.66
## 4 AKLAN                     7.31        8.65      8.79        6.68         8.12
## 5 ALBAY                     7.03        8.99      8.90        7.03         7.27
## 6 ANTIQUE                   7.72        8.56      9.23        7.88         7.58
## 7 APAYAO                    7.73        7.94      8.30        7.53         2.80
## 8 AURORA                    7.94        9.21      8.66        7.46         7.48
## # i 6 more variables: AVG_NET_FOR <dbl>, AVG_NET_RET <dbl>, AVG_NET_COM <dbl>,
## #   AVG_NET_CONS <dbl>, AVG_NET_NEC <dbl>, AVG_HHI <dbl>
```

```r
#Identify common provinces

# Identify provinces present in both datasets
common_provinces <- intersect(fies_aggregated$Province, ph$Province)

# Filter both datasets to retain only common provinces
FIES_HHI <- fies_aggregated %>%
  filter(Province %in% common_provinces)

SHP_HHI <- ph %>%
  filter(Province %in% common_provinces) %>%
  arrange(Province)

# Check dimensions of the filtered datasets
dim(FIES_HHI)
```

```
## [1] 82 12
```

```r
dim(SHP_HHI)
```

```
## [1] 82  9
```

```r
FIES_HHI_SHP <- left_join(FIES_HHI, SHP_HHI, by = "Province")
colnames(FIES_HHI_SHP)
```

```
##  [1] "Province"     "AVG_NETSHARE" "AVG_PENSION"  "AVG_REGFT"    "AVG_NET_LPR"
##  [6] "AVG_NET_FISH" "AVG_NET_FOR"  "AVG_NET_RET"  "AVG_NET_COM"  "AVG_NET_CONS"
## [11] "AVG_NET_NEC"  "AVG_HHI"      "X2004"        "X2007"        "X2010"
## [16] "X2013"        "X2016"        "lon"          "lat"          "geometry"
```

```r
head(FIES_HHI_SHP)
```

```
## # A tibble: 6 x 20
##   Province          AVG_NETSHARE AVG_PENSION AVG_REGFT AVG_NET_LPR AVG_NET_FISH
##   <chr>                    <dbl>       <dbl>     <dbl>       <dbl>        <dbl>
## 1 ABRA                      7.95        8.66      8.37        8.03         5.95
## 2 AGUSAN DEL NORTE          7.13        9.22      8.39        7.28         6.62
## 3 AGUSAN DEL SUR            7.05        8.73      8.10        7.30         4.66
## 4 AKLAN                     7.31        8.65      8.79        6.68         8.12
## 5 ALBAY                     7.03        8.99      8.90        7.03         7.27
## 6 ANTIQUE                   7.72        8.56      9.23        7.88         7.58
## # i 14 more variables: AVG_NET_FOR <dbl>, AVG_NET_RET <dbl>, AVG_NET_COM <dbl>,
## #   AVG_NET_CONS <dbl>, AVG_NET_NEC <dbl>, AVG_HHI <dbl>, X2004 <dbl>,
## #   X2007 <dbl>, X2010 <dbl>, X2013 <dbl>, X2016 <dbl>, lon <dbl>, lat <dbl>,
## #   geometry <MULTIPOLYGON [°]>
```

```r
head(FIES_HHI)
```

```
## # A tibble: 6 x 12
##   Province          AVG_NETSHARE AVG_PENSION AVG_REGFT AVG_NET_LPR AVG_NET_FISH
##   <chr>                    <dbl>       <dbl>     <dbl>       <dbl>        <dbl>
## 1 ABRA                      7.95        8.66      8.37        8.03         5.95
## 2 AGUSAN DEL NORTE          7.13        9.22      8.39        7.28         6.62
## 3 AGUSAN DEL SUR            7.05        8.73      8.10        7.30         4.66
## 4 AKLAN                     7.31        8.65      8.79        6.68         8.12
## 5 ALBAY                     7.03        8.99      8.90        7.03         7.27
## 6 ANTIQUE                   7.72        8.56      9.23        7.88         7.58
## # i 6 more variables: AVG_NET_FOR <dbl>, AVG_NET_RET <dbl>, AVG_NET_COM <dbl>,
## #   AVG_NET_CONS <dbl>, AVG_NET_NEC <dbl>, AVG_HHI <dbl>
```

```r
summary(FIES_HHI)
```

```
##    Province          AVG_NETSHARE    AVG_PENSION      AVG_REGFT
##  Length:82          Min.   :3.972   Min.   :1.947   Min.   :5.895
##  Class :character   1st Qu.:6.908   1st Qu.:8.174   1st Qu.:8.369
##  Mode  :character   Median :7.410   Median :8.632   Median :8.620
##                     Mean   :7.093   Mean   :8.475   Mean   :8.540
##                     3rd Qu.:7.695   3rd Qu.:8.979   3rd Qu.:8.833
##                     Max.   :8.228   Max.   :9.431   Max.   :9.363
##   AVG_NET_LPR     AVG_NET_FISH     AVG_NET_FOR     AVG_NET_RET
##  Min.   :1.317   Min.   : 1.332   Min.   :0.000   Min.   : 7.235
##  1st Qu.:7.035   1st Qu.: 6.706   1st Qu.:4.792   1st Qu.: 9.260
##  Median :7.347   Median : 7.591   Median :5.524   Median : 9.509
##  Mean   :7.092   Mean   : 7.112   Mean   :5.068   Mean   : 9.504
##  3rd Qu.:7.739   3rd Qu.: 8.295   3rd Qu.:5.940   3rd Qu.: 9.750
##  Max.   :9.181   Max.   :10.956   Max.   :7.863   Max.   :10.427
##   AVG_NET_COM     AVG_NET_CONS    AVG_NET_NEC       AVG_HHI
##  Min.   :0.000   Min.   :0.000   Min.   :1.459   Min.   :3.306
##  1st Qu.:5.734   1st Qu.:1.647   1st Qu.:5.907   1st Qu.:4.081
##  Median :6.155   Median :2.884   Median :6.742   Median :4.328
##  Mean   :5.945   Mean   :2.972   Mean   :6.364   Mean   :4.382
##  3rd Qu.:6.514   3rd Qu.:4.613   3rd Qu.:7.306   3rd Qu.:4.700
##  Max.   :7.628   Max.   :6.450   Max.   :9.100   Max.   :5.665
```

```r
colnames(FIES_HHI)
```

```
##  [1] "Province"     "AVG_NETSHARE" "AVG_PENSION"  "AVG_REGFT"    "AVG_NET_LPR"
##  [6] "AVG_NET_FISH" "AVG_NET_FOR"  "AVG_NET_RET"  "AVG_NET_COM"  "AVG_NET_CONS"
## [11] "AVG_NET_NEC"  "AVG_HHI"
```

```r
OLS_model = lm(AVG_HHI~. -Province, data = FIES_HHI)
summary(OLS_model)
```

```
##
## Call:
## lm(formula = AVG_HHI ~ . - Province, data = FIES_HHI)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79848 -0.29519  0.01726  0.28527  0.59060
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.86270    1.95585   1.464 0.147700
## AVG_NETSHARE -0.06499    0.06450  -1.008 0.317062
## AVG_PENSION  -0.08408    0.07678  -1.095 0.277175
## AVG_REGFT     0.54147    0.15125   3.580 0.000625 ***
## AVG_NET_LPR  -0.15142    0.06804  -2.225 0.029229 *
## AVG_NET_FISH -0.04645    0.02745  -1.692 0.094998 .
## AVG_NET_FOR  -0.04347    0.03713  -1.171 0.245581
## AVG_NET_RET   0.14663    0.21465   0.683 0.496743
## AVG_NET_COM  -0.10239    0.07761  -1.319 0.191315
## AVG_NET_CONS -0.07036    0.03170  -2.219 0.029656 *
## AVG_NET_NEC  -0.13871    0.04241  -3.270 0.001658 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3836 on 71 degrees of freedom
## Multiple R-squared:  0.4925, Adjusted R-squared:  0.421
## F-statistic:  6.89 on 10 and 71 DF,  p-value: 1.999e-07
```

```r
head(ph)
```

```
## Simple feature collection with 6 features and 8 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: 118.0642 ymin: 4.587294 xmax: 125.9392 ymax: 8.163486
## Geodetic CRS:  WGS 84
##          Province     X2004     X2007     X2010     X2013     X2016      lon
## 1   LANAO DEL SUR  46.84711  49.89349  68.73035  74.60973  67.80187 124.3362
## 2     MAGUINDANAO 126.93136 150.12959 152.62346 115.27281 102.51912 124.3847
## 3            SULU  85.17795  91.14583  87.68201 141.89609  96.14512 121.0543
## 4         BASILAN 137.90100 144.79500  90.66358  97.50297 121.16571 122.0291
## 5       TAWI-TAWI 109.56903 109.09091 133.92857 131.82160 153.89351 119.9010
## 6 DAVAO DEL NORTE  88.68584  88.38384  90.67952  90.67952  91.82736 125.6423
##        lat                       geometry
```

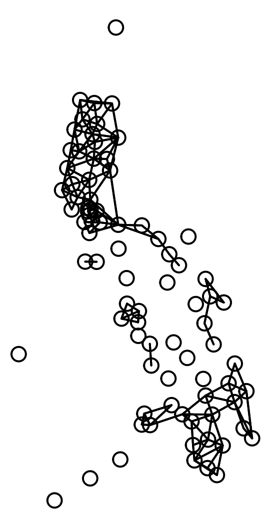
```
## 1 7.796685 MULTIPOLYGON (((124.484 8.1...
## 2 7.031549 MULTIPOLYGON (((124.047 7.3...
## 3 5.953217 MULTIPOLYGON (((120.6989 6....
## 4 6.565456 MULTIPOLYGON (((121.5931 6....
## 5 5.239002 MULTIPOLYGON (((118.4204 7....
## 6 7.585208 MULTIPOLYGON (((125.3794 8....
```

```r
# create list of neighbors, based on contiguity, from the Philippine spatial data
PHneighbor <- poly2nb(SHP_HHI)
```

```
## Warning in poly2nb(SHP_HHI): some observations have no neighbours;
## if this seems unexpected, try increasing the snap argument.
```

```
## Warning in poly2nb(SHP_HHI): neighbour object has 21 sub-graphs;
## if this sub-graph count seems unexpected, try increasing the snap argument.
```

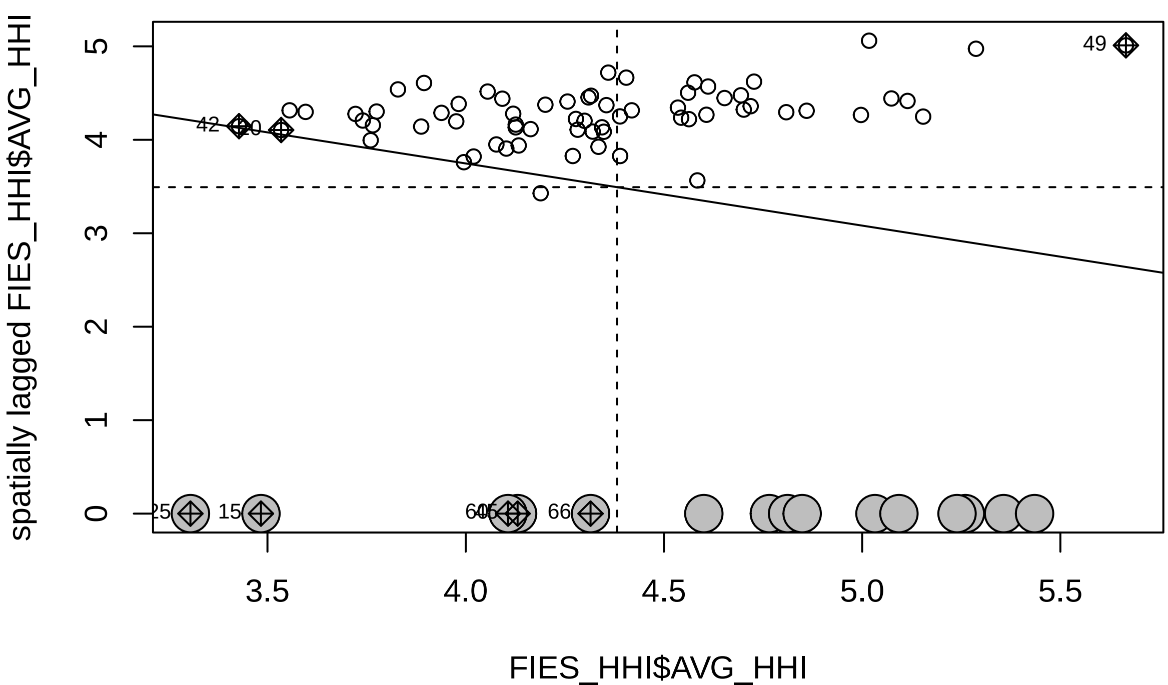
```r
# create list of centroid coordinates from the Philippine spatial data
geometry <- st_geometry(SHP_HHI)
PHcoords <- st_centroid(geometry)

# visualize the connections between Philippine provinces
plot(PHneighbor, PHcoords)
```



```r
#Create contiguity
PH_list <- nb2listw(PHneighbor, zero.policy = TRUE)

#Do spatial autocorrelation
moran.test(FIES_HHI$AVG_HHI, PH_list)
```

```
##
##  Moran I test under randomisation
##
## data:  FIES_HHI$AVG_HHI
## weights: PH_list
## n reduced by no-neighbour observations
##
## Moran I statistic standard deviate = 2.1678, p-value = 0.01509
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##       0.198679015      -0.015151515       0.009730103
```

```r
moran.plot(FIES_HHI$AVG_HHI, PH_list)
```




```r
# Lagrange multiplier test for spatial lag and spatial error dependencies
lm.LMtests(OLS_model, PH_list, test=c("LMlag", "LMerr"))
```

```
## Please update scripts to use lm.RStests in place of lm.LMtests
```

```
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
##  dependence
##
## data:
## model: lm(formula = AVG_HHI ~ . - Province, data = FIES_HHI)
## test weights: listw
##
## RSlag = 0.056866, df = 1, p-value = 0.8115
##
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
##  dependence
##
## data:
## model: lm(formula = AVG_HHI ~ . - Province, data = FIES_HHI)
## test weights: listw
##
## RSerr = 6.0215, df = 1, p-value = 0.01413
```

Under contiguity, the error model is significant.

```r
# Spatial error model
# Example of creating the lagged variable using the spatial weights matrix

spatial_error <- errorsarlm(AVG_HHI ~. -Province, data = FIES_HHI, listw = PH_list, zero.policy= TRUE)
summary(spatial_error)
```

```
##
## Call:errorsarlm(formula = AVG_HHI ~ . - Province, data = FIES_HHI,
##     listw = PH_list, zero.policy = TRUE)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.777705 -0.295775  0.051076  0.307760  0.497642
##
## Type: error
## Regions with no neighbours included:
##  9 11 14 15 21 23 25 31 44 45 60 66 69 74 78
## Coefficients: (asymptotic standard errors)
##               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)   2.548027   1.720531  1.4810  0.138619
## AVG_NETSHARE -0.043346   0.064830 -0.6686  0.503747
## AVG_PENSION  -0.065883   0.069247 -0.9514  0.341392
## AVG_REGFT     0.567360   0.141386  4.0128 5.999e-05
## AVG_NET_LPR  -0.130842   0.064363 -2.0329  0.042065
## AVG_NET_FISH -0.028840   0.025407 -1.1351  0.256335
## AVG_NET_FOR  -0.067490   0.033963 -1.9872  0.046906
## AVG_NET_RET   0.114191   0.188936  0.6044  0.545585
## AVG_NET_COM  -0.097864   0.067721 -1.4451  0.148429
## AVG_NET_CONS -0.071487   0.027203 -2.6279  0.008591
## AVG_NET_NEC  -0.148342   0.036474 -4.0671 4.760e-05
##
```

```
## Lambda: 0.36256, LR test value: 6.6947, p-value: 0.00967
## Asymptotic standard error: 0.12003
##     z-value: 3.0205, p-value: 0.0025233
## Wald statistic: 9.1236, p-value: 0.0025233
##
## Log likelihood: -28.53072 for error model
## ML residual variance (sigma squared): 0.11294, (sigma: 0.33607)
## Number of observations: 82
## Number of parameters estimated: 13
## AIC: 83.061, (AIC for lm: 87.756)
```

```r
stargazer(OLS_model, spatial_error, type = "text",
                    title="Regression Results from OLS and with Spatial Error")
```

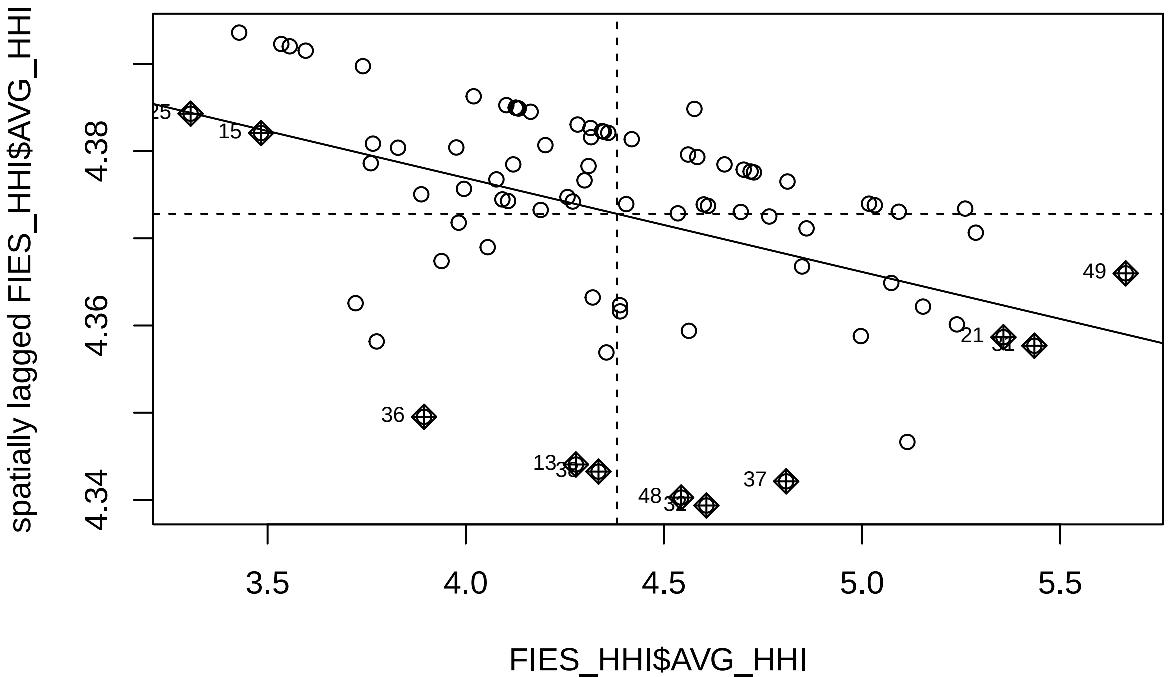
```
##
## Regression Results from OLS and with Spatial Error
## ============================================================
##                                 Dependent variable:
##                     ----------------------------------------
##                                       AVG_HHI
##                               OLS                spatial
##                                                   error
##                               (1)                  (2)
## ------------------------------------------------------------
## AVG_NETSHARE                -0.065               -0.043
##                             (0.065)              (0.065)
##
## AVG_PENSION                 -0.084               -0.066
##                             (0.077)              (0.069)
##
## AVG_REGFT                   0.541***             0.567***
##                             (0.151)              (0.141)
##
## AVG_NET_LPR                 -0.151**             -0.131**
##                             (0.068)              (0.064)
##
## AVG_NET_FISH                -0.046*              -0.029
##                             (0.027)              (0.025)
##
## AVG_NET_FOR                 -0.043               -0.067**
##                             (0.037)              (0.034)
##
## AVG_NET_RET                  0.147                0.114
##                             (0.215)              (0.189)
##
## AVG_NET_COM                 -0.102               -0.098
##                             (0.078)              (0.068)
##
## AVG_NET_CONS                -0.070**             -0.071***
##                             (0.032)              (0.027)
##
## AVG_NET_NEC                -0.139***             -0.148***
##                             (0.042)              (0.036)
##
```

```
## Constant                       2.863                2.548
##                               (1.956)              (1.721)
##
## ------------------------------------------------------------
## Observations                     82                   82
## R2                             0.493
## Adjusted R2                    0.421
## Log Likelihood                                     -28.531
## sigma2                                               0.113
## Akaike Inf. Crit.                                   83.061
## Residual Std. Error    0.384 (df = 71)
## F Statistic         6.890*** (df = 10; 71)
## Wald Test                                 9.124*** (df = 1)
## LR Test                                   6.695*** (df = 1)
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

```
lat_long = st_drop_geometry(SHP_HHI)[,c("lat", "lon")]
lat_long = as.matrix(lat_long)
PHneighbor_D = dnearneigh(lat_long, d1=0, d2=10)
PHlistw_D = nb2listw(PHneighbor_D, style="W")

# test for spatial dependence using Moran's I test
moran.test(FIES_HHI$AVG_HHI, PHlistw_D)
```

```
##
##  Moran I test under randomisation
##
## data:  FIES_HHI$AVG_HHI
## weights: PHlistw_D
##
## Moran I statistic standard deviate = 0.32945, p-value = 0.3709
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -1.077295e-02     -1.234568e-02      2.278893e-05
```

```
moran.plot(FIES_HHI$AVG_HHI, PHlistw_D)
```




```r
# Lagrange multiplier test for spatial lag and spatial error dependencies
lm.LMtests(OLS_model, PHlistw_D, test=c("LMlag", "LMerr"))
```

```
## Please update scripts to use lm.RStests in place of lm.LMtests

##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
##  dependence
##
## data:
## model: lm(formula = AVG_HHI ~ . - Province, data = FIES_HHI)
## test weights: listw
##
## RSlag = 0.26696, df = 1, p-value = 0.6054
##
##
##  Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial
##  dependence
##
## data:
## model: lm(formula = AVG_HHI ~ . - Province, data = FIES_HHI)
## test weights: listw
##
## RSerr = 0.012318, df = 1, p-value = 0.9116
```

Using distance matrix is not significant. We should stick to contiguity and error only.