

Countries_script

2025-06-13

To start we read in the data and put the tidyverse package in the library. Remember you must install this package once and only once on your machine.

```
# install.packages("tidyverse")
library(tidyverse)
# install.packages("RColorBrewer")
library(RColorBrewer)
# install.packages("scales")
library(scales)
```

```
gdp_df <- read.csv("GDP.csv")
Inf_mort_df <- read.csv("Inf_mort.csv")
Life_exp_df <- read.csv("Life_exp.csv")
Pop_df <- read.csv("Pop.csv")

View(gdp_df)
View(Inf_mort_df)
View(Life_exp_df)
View(Pop_df)
```

Next we merge the four data files using the merge function.

```
country_df <- merge(Pop_df, Life_exp_df, by = "Country")
country_df <- merge(country_df, Inf_mort_df, by = "Country")
country_df <- merge(country_df, gdp_df, by = "Country")

View(country_df)
```

The histograms can be plotted for all four data files by setting up a two-by-two matrix.

```
# set parameter to 2x2 window
par(mfrow = c(2,2))

hist(country_df$GDP_cap, col = "grey", main = "GDP per Capita", xlab="GDP_per_cap")
abline(v = mean(country_df$GDP_cap), col="blue")
abline(v = median(country_df$GDP_cap), col="red", lty = 2)

hist(country_df$Population[country_df$Population < 1*10^9], col = "grey",
      main = "Population", xlab="People")
abline(v = mean(country_df$Population), col="blue")
abline(v = median(country_df$Population), col="red", lty = 2)
```

```

hist(country_df$Inf_mort, col = "grey",main = "Infant Mortality per 1,000",
      xlab="Inf Mort rate")
abline(v = mean(country_df$Inf_mort), col="blue")
abline(v = median(country_df$Inf_mort), col="red", lty = 2)

hist(country_df$Life_exp, col = "grey",main = "Life Expectancy", xlab="Age")
abline(v = mean(country_df$Life_exp), col="blue")
abline(v = median(country_df$Life_exp), col="red", lty = 2)

```

The scatterplot was first created using the base R plot function but required some work to correctly size the bubbles by population using deciles.

```

#Basic scatterplot using plot function in Base R package
# pch sets type of point get help on pch to see options
plot(country_df$GDP_cap, country_df$Life_exp, pch = 19,
      col = grey(0.05, alpha = 0.5),
      xlab = "GDP_per_cap", ylab = "Life expectancy")

?pch

#Scatterplot try to size bubbles by Population using cex

plot(country_df$GDP_cap, country_df$Life_exp, pch = 19,
      col = grey(0.05, alpha = 0.5),
      cex = country_df$Population,
      xlab = "GDP_per_cap", ylab = "Life expectancy")

#Yikes! cex uses raw value of vector and Populations are HUGE!
#We will first compute deciles using quantile function
bins = quantile(country_df$Population, prob = (0:10)/10)
bins

#Note bins is a vector of length 11
#Next we assign each country its decile 1-10 using the cut function
#Note vals is a vector of length 219 same as Population vector
vals = cut(country_df$Population, bins, labels = F, include.lowest = T)
vals

#Now we can size circles by the vals
plot(country_df$GDP_cap, country_df$Life_exp, pch = 19,
      col = grey(0.05, alpha = 0.5),
      cex = 0.25*vals,
      xlab = "GDP_per_cap", ylab = "Life expectancy")

#We will add a best fit trendline using the lm function
#y is a function of x so we type y ~ x
lm(country_df$Life_exp ~ country_df$GDP_cap)

#now use the abline function as for the histograms

?abline

abline(v=median(country_df$GDP_cap), col="red", lty = 2)

```

```
abline(h=median(country_df$Life_exp), col="red", lty = 2)
abline(lm(country_df$Life_exp ~ country_df$GDP_cap), col = "blue", lwd = 2)
```

Next we create the scatterplot using the ggplot function.

```
ggplot(country_df, aes(GDP_cap, Life_exp, size = Population,
                      fill = Inf_mort)) +
  geom_point(shape = 21, color = "black",
            alpha = 0.5) +
  scale_size_area(max_size = 15) +
  labs(x = "GDP per Capita",
       y = "Life Expectancy in years") +
  scale_fill_gradient(low = "blue", high = "red") +
  scale_x_continuous(labels = label_dollar())
```

Boxplots require us to create the wealth vector using logical subsetting.

```
# First create Wealth vector that repeats "Poor"
country_df$Wealth = rep("Poor", nrow(country_df))

# Next overwrite all Wealth values with "Rich" satisfying criterion
country_df$Wealth[country_df$GDP_cap > 60000] = "Rich"
head(country_df)

boxplot(country_df$Life_exp, horizontal = T, col = "purple",
        xlab = "Life expectancy")

qplot(Wealth, Life_exp, data = country_df, geom = "boxplot",
      fill = Wealth)

ggplot(country_df, aes(Wealth, Life_exp)) +
  geom_boxplot(aes(fill = Wealth), show.legend = FALSE) +
  ylab("Years") + labs(title = "Life Expectancy")
```

Finally to get the heat map required a couple new packages and merging files.

```
#install.packages("maps")
library(maps)
#install.packages("mapproj")
library(mapproj)

world_map <- map_data("world")
ggplot(world_map, aes(long, lat, group = group)) +
  geom_polygon(fill = "white", color = "blue") +
  coord_quickmap()

# Easier to rename just the rows in country_df for USA and UK
which(country_df$Country == "United States")
country_df$Country[208] <- "UK"
country_df$Country[209] <- "USA"

#First let's add region vector to country_df that is same as Country variable
```

```

country_df$region <- country_df$Country
head(country_df)
head(world_map)

##let's try merging using all.x = TRUE
choro_df = merge(world_map, country_df, by = "region", all.x = TRUE)

#Aaargh the order variable got messed up in the merge
# So we need to reorder by order
choro <- choro_df[order(choro_df$order), ]

# We can finally get the heat map
ggplot(choro, aes(long, lat, group = group, fill = IM_rank)) +
  geom_polygon(color = "black") +
  scale_fill_gradient(low = "red", high = "yellow")

```

The End.