Data preparation

setwd("D:\\temp")

library(XLConnect)

wk <- loadWorkbook("GCI\_dataset.xlsx")

df <- readWorksheet(wk, sheet="Sheet1")

#fix PHL data issue

df$PHL = as.numeric(df$PHL)

**Q1**

library(dplyr)

hiv\_prevalence <- df %>% filter(Edition=="2014-2015", Series.code=="4.05", Attribute=="Value") %>% select(KHM:VNM)

colnames <- colnames(hiv\_prevalence)

hiv\_prevalence <- t(hiv\_prevalence)

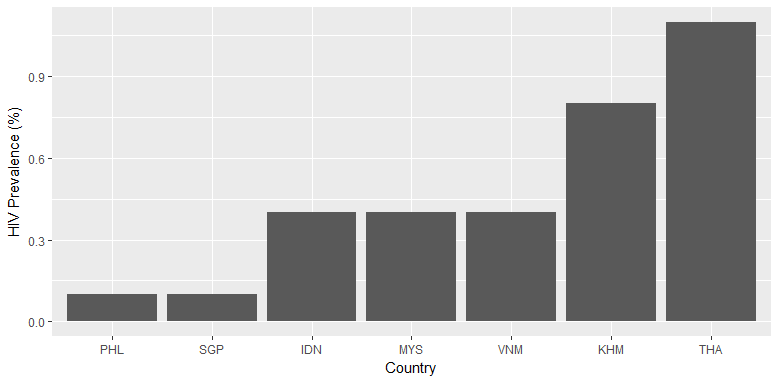
hiv\_df = data.frame(colnames, hiv\_prevalence)

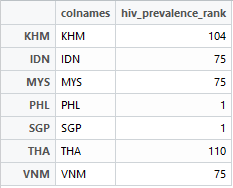
library(ggplot2)

ggplot(hiv\_df, aes(x=reorder(colnames, hiv\_prevalence), weight=hiv\_prevalence)) +

geom\_bar() +

labs(x="Country", y="HIV Prevalence (%)")





hiv\_prevalence\_rank <- df %>% filter(Edition=="2014-2015", Series.code=="4.05", Attribute=="Rank") %>% select(KHM:VNM)

colnames <- colnames(hiv\_prevalence\_rank)

hiv\_prevalence\_rank <- t(hiv\_prevalence\_rank)

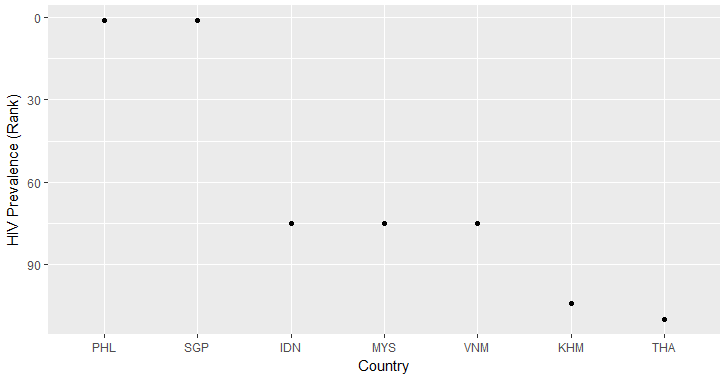
hiv\_df\_rank <- data.frame(colnames, hiv\_prevalence\_rank)

ggplot(hiv\_df\_rank, aes(x=reorder(colnames, hiv\_prevalence\_rank), y=hiv\_prevalence\_rank)) +

geom\_point() +

labs(x="Country", y="HIV Prevalence (Rank)") +

scale\_y\_reverse()



From the plot, among the 7 countries, Philippines and Singapore are the highest rank (first rank). Followed by Indonesia, Malaysia, and Vietnam, all have the same rank, ranked 75th. Cambodia is ranked 104th and Thailand has the worst HIV prevalence among the 7 countries, ranked 110th.

**Q2**

irreg\_pay\_cambodia = df %>% filter(Series.code=="1.05", Attribute=="Value") %>% select(Edition, KHM)

reliability\_police\_cambodia = df %>% filter(Series.code=="1.16", Attribute=="Value") %>% select(Edition, KHM)

cor(irreg\_pay\_cambodia$KHM, reliability\_police\_cambodia$KHM[1:5])

# [1] 0.8921799

With the correlation of 0.8921799, ‘Irregular payments and bribes’ and ‘Reliability of Police Services’ in Cambodia are highly correlated. The correlation is visualized below:

ggplot(subset(df, Series.code %in% c("1.05", "1.16") & Attribute=="Value"),

aes(x=substr(Edition,1,4),

y=KHM,

group=Series.code,

color=factor(Series.code))) +

geom\_point() +

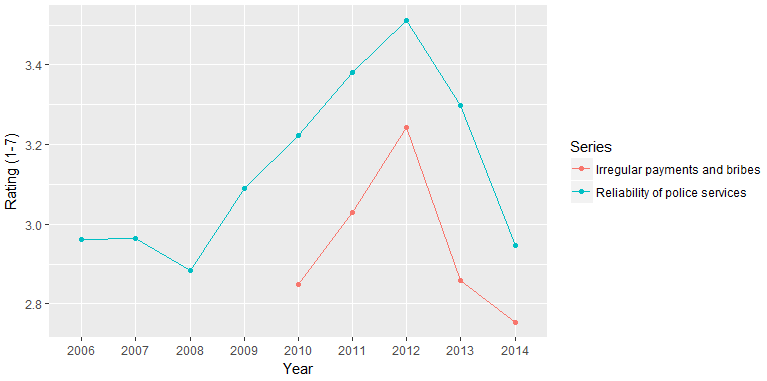
geom\_line() +

labs(x="Year", y="Rating (1-7)") +

scale\_color\_discrete(name="Series",

breaks=c("1.05", "1.16"),

labels=c("Irregular payments and bribes", "Reliability of police services"))



library(tidyr)

df2 <- df1[names(df1) != "Series"]

df2$Series.code <- paste("c", df2$Series.code, sep="")

df2 <- spread(df2[names(df2) != "Series"], Series.code, Value)

ggplot(df2[df2$Attribute=="Value" & df2$Country=="KHM",],

aes(x=c1.05,

y=c1.16,

group=1,

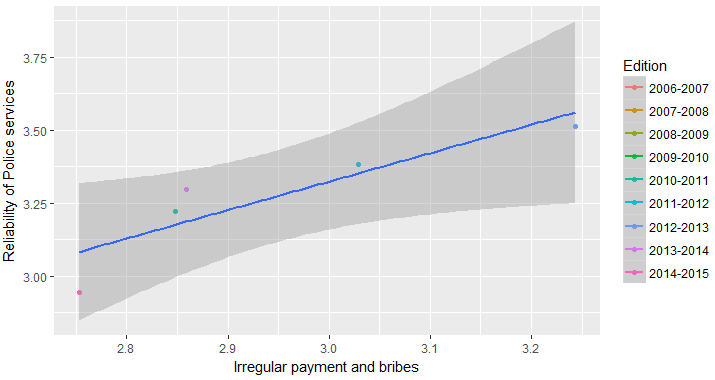
color=Edition))+

geom\_point()+

geom\_smooth(method="lm")+

labs(x="Irregular payment and bribes",

y="Reliability of Police services")



**Q3**

df1 <- df %>% gather(Country, Value, KHM:VNM)

ggplot(df1[df1$Series.code=="10.01" & df1$Attribute=="Value",],

aes(x=substr(Edition,1,4),

y=Value,

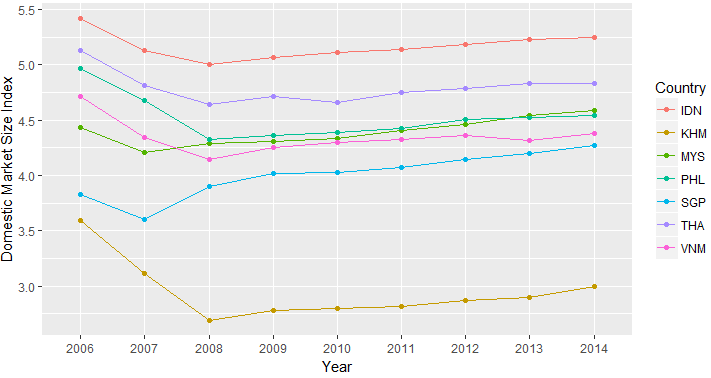
group=Country,

color=Country)) +

geom\_point() +

geom\_line() +

labs(x="Year", y="Domestic Market Size Index")



ggplot(df1[df1$Series.code=="10.02" & df1$Attribute=="Value",],

aes(x=substr(Edition,1,4),

y=Value,

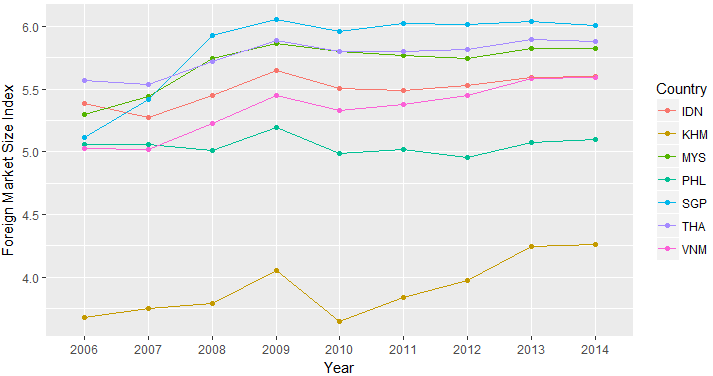
group=Country,

color=Country)) +

geom\_point() +

geom\_line() +

labs(x="Year", y="Foreign Market Size Index")



From the previous 2 charts, it seems that the index tends to move at the same direction. However, the “index” may not be comparable year-on-year. Based on the definition of the index which is normalized from range 1 to 7, it is highly dependent on the country having the highest market size (which will be at scale 7), where the market size of other countries will be scaled accordingly.

ggplot(df1[df1$Series.code=="10.01" & df1$Attribute=="Rank",],

aes(x=substr(Edition,1,4),

y=Value,

group=Country,

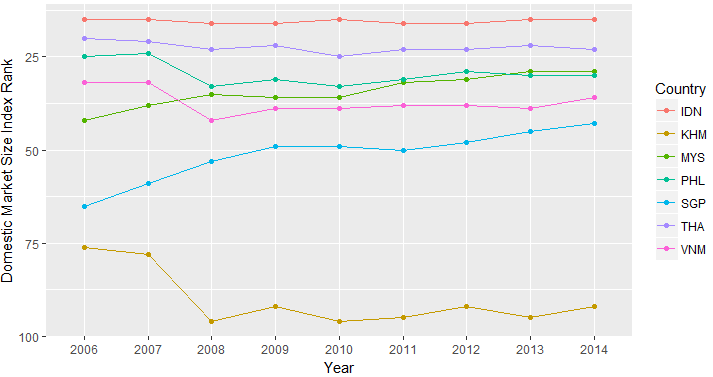
color=Country)) +

geom\_point() +

geom\_line() +

labs(x="Year", y="Domestic Market Size Index Rank") +

scale\_y\_reverse()



From the plot, it seems that Singapore and Malaysia have a trend of increasing in Domestic Market Size Rank. The rest of the countries show smaller increase or fluctuating around the same ranks.

ggplot(df1[df1$Series.code=="10.02" & df1$Attribute=="Rank",],

aes(x=substr(Edition,1,4),

y=Value,

group=Country,

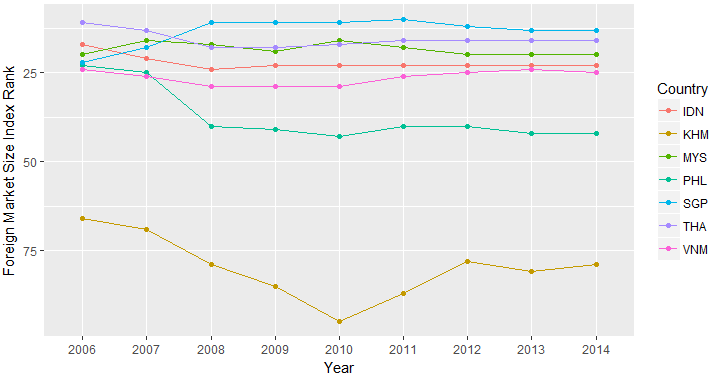
color=Country)) +

geom\_point() +

geom\_line() +

labs(x="Year", y="Foreign Market Size Index Rank") +

scale\_y\_reverse()



From the plot, it seems that there was a significant change in ranks around year 2008, and during that time Singapore reached the highest foreign market size among the 7 countries and keep the highest position. Cambodia are having the smallest domestic and foreign market size among the 7 countries.

**Q4**

# clean data

df1$Value <- df1$Value \* 10^(-3\*(df1$Value > 100000 & df1$Series.code=="0.01" & df1$Attribute=="Value"))

ggplot(df1[df1$Series.code=="0.01" & df1$Attribute=="Value" & (df1$Country=="IDN"|df1$Country=="THA"),],

aes(x=substr(Edition,1,4),

y=Value,

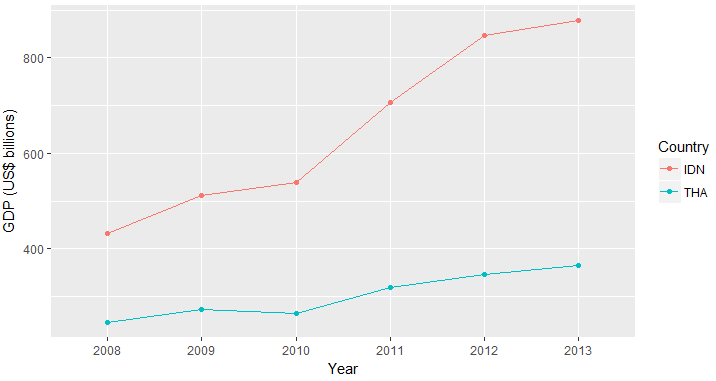
group=Country,

color=Country))+

geom\_point()+

geom\_line()+

labs(x="Year", y="GDP (US$ billions)")



cor(

df1[df1$Series.code=="0.01" & df1$Attribute=="Value" & df1$Country=="IDN",]$Value,

df1[df1$Series.code=="0.01" & df1$Attribute=="Value" & df1$Country=="THA",]$Value

)

# [1] 0.9908999

It seems that the GDP of Indonesia and Thailand over the years are highly correlated

ggplot(df1[df1$Series.code=="5.03" & df1$Attribute=="Value" & (df1$Country=="IDN"|df1$Country=="THA"),],

aes(x=substr(Edition,1,4),

y=Value,

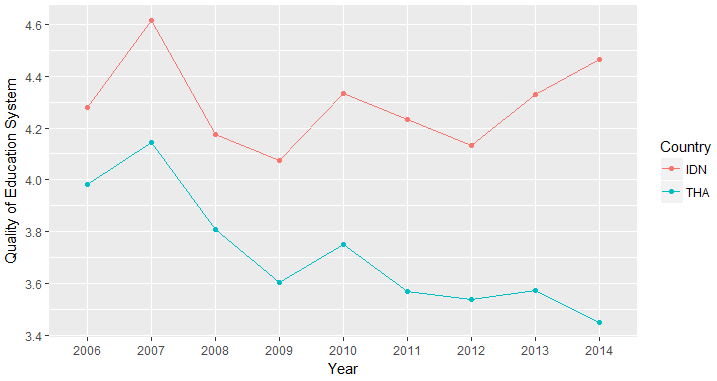
group=Country,

color=Country))+

geom\_point()+

geom\_line()+

labs(x="Year", y="Quality of Education System")



cor(

df1[df1$Series.code=="5.03" & df1$Attribute=="Value" & df1$Country=="IDN",]$Value,

df1[df1$Series.code=="5.03" & df1$Attribute=="Value" & df1$Country=="THA",]$Value

)

# [1] 0.4382189

The quality of the education systems between Indonesia and Thailand are moderately correlated.

df2 <- df1[names(df1) != "Series"]

df2$Series.code <- paste("c", df2$Series.code, sep="")

df2 <- spread(df2[names(df2) != "Series"], Series.code, Value)

cor(

df1[df1$Series.code=="0.01" & df1$Attribute=="Value" & df1$Country=="IDN" & as.integer(substr(df1$Edition, 1, 4)) >= 2008 & as.integer(substr(df1$Edition, 1, 4)) <= 2013,]$Value,

df1[df1$Series.code=="5.03" & df1$Attribute=="Value" & df1$Country=="IDN" & as.integer(substr(df1$Edition, 1, 4)) >= 2008 & as.integer(substr(df1$Edition, 1, 4)) <= 2013,]$Value

)

# [1] 0.2664105

In Indonesia, there is a slight positive correlation between the quality of the education systems and GDP.

ggplot(df2[df2$Attribute=="Value" & df2$Country=="IDN" & as.integer(substr(df2$Edition, 1, 4)) >= 2008 & as.integer(substr(df2$Edition, 1, 4)) <= 2013,],

aes(x=c5.03,

y=c0.01,

group=1,

color=Edition))+

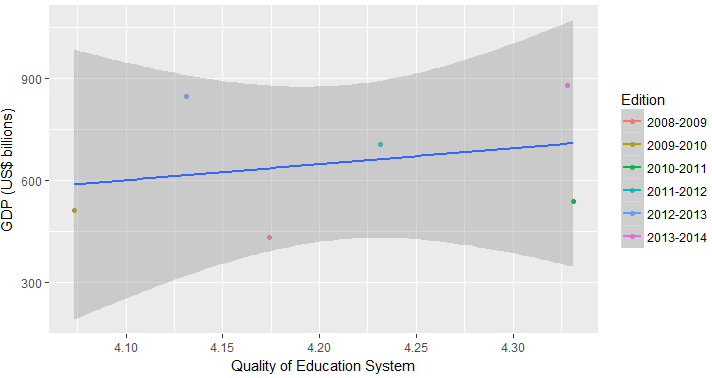
geom\_point()+

geom\_smooth(method="lm")+

labs(x="Quality of Education System",

y="GDP (US$ billions)")

Indonesia:



cor(

df1[df1$Series.code=="0.01" & df1$Attribute=="Value" & df1$Country=="THA" & as.integer(substr(df1$Edition, 1, 4)) >= 2008 & as.integer(substr(df1$Edition, 1, 4)) <= 2013,]$Value,

df1[df1$Series.code=="5.03" & df1$Attribute=="Value" & df1$Country=="THA" & as.integer(substr(df1$Edition, 1, 4)) >= 2008 & as.integer(substr(df1$Edition, 1, 4)) <= 2013,]$Value

)

# [1] -0.8402767

In Thailand however, there is a negative correlation between quality of the education system and GDP.

ggplot(df2[df2$Attribute=="Value" & df2$Country=="THA" & as.integer(substr(df2$Edition, 1, 4)) >= 2008 & as.integer(substr(df2$Edition, 1, 4)) <= 2013,],

aes(x=c5.03,

y=c0.01,

group=1,

color=Edition))+

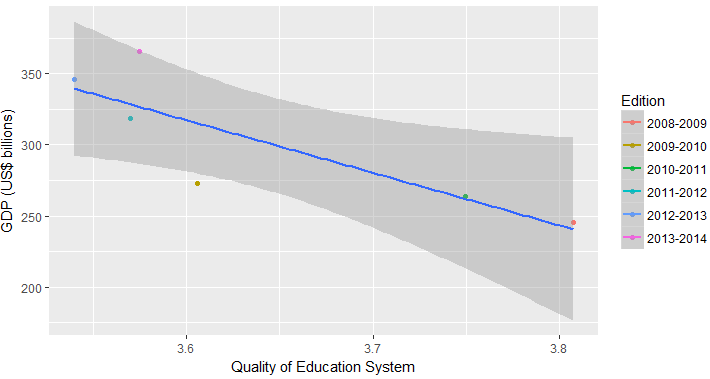
geom\_point()+

geom\_smooth(method="lm")+

labs(x="Quality of Education System",

y="GDP (US$ billions)")

Thailand:



cor(

df1[df1$Series.code=="0.01" & df1$Attribute=="Value" & (df1$Country=="IDN" | df1$Country=="THA") & as.integer(substr(df1$Edition, 1, 4)) >= 2008 & as.integer(substr(df1$Edition, 1, 4)) <= 2013,]$Value,

df1[df1$Series.code=="5.03" & df1$Attribute=="Value" & (df1$Country=="IDN" | df1$Country=="THA") & as.integer(substr(df1$Edition, 1, 4)) >= 2008 & as.integer(substr(df1$Edition, 1, 4)) <= 2013,]$Value

)

# [1] 0.7769784

ggplot(df2[df2$Attribute=="Value" & (df2$Country=="IDN" | df2$Country=="THA"),],

aes(x=c5.03,

y=c0.01,

group=1,

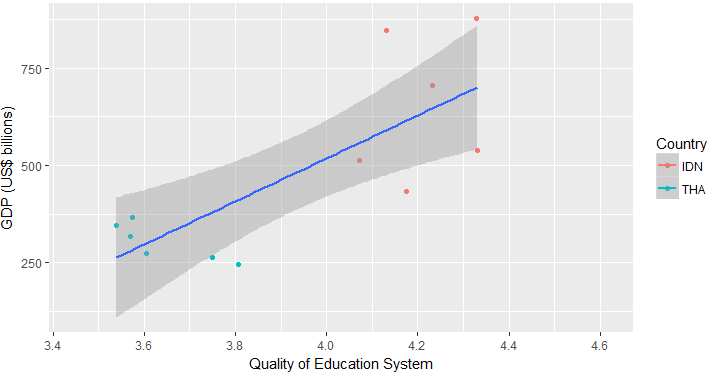
color=Country))+

geom\_point()+

geom\_smooth(method="lm")+

labs(x="Quality of Education System",

y="GDP (US$ billions)")



When we combine the data for both Indonesia and Thailand, generally Thailand has lower quality of education system and lower GDP as compared to Indonesia, and there is a positive correlation between Quality of the Education System and GDP. The higher the Quality of the Education System, the higher the GDP.

**Q5**

df2 <- df1[names(df1) != "Series"]

df2$Series.code <- paste("c", df2$Series.code, sep="")

df2 <- spread(df2[names(df2) != "Series"], Series.code, Value)

ggplot(df2[df2$Attribute=="Value",],

aes(x=c3.03,

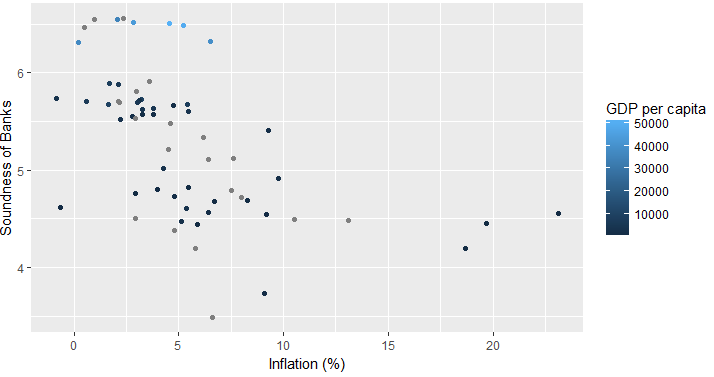
y=c8.06,

group=Country,

color=c0.03))+

geom\_point()+

labs(x="Inflation (%)", y="Soundness of Banks", color="GDP per capita")



From the plot, generally the one with high GDP per capita are on the top (having high Soundness of Banks). However, it seems that there is little correlation between inflation and GDP per capita. There are quite a number of data points having low inflation with various level of GDP per capita.

data.matrix(cor(df2[df2$Attribute=="Value" & as.integer(substr(df2$Edition, 1, 4)) >= 2008 & as.integer(substr(df2$Edition, 1, 4)) <= 2013,] %>% select(c0.03, c3.03, c8.06)))

# c0.03 c3.03 c8.06

# c0.03 1.0000000 -0.2090328 0.7293571

# c3.03 -0.2090328 1.0000000 -0.5214003

# c8.06 0.7293571 -0.5214003 1.0000000

As shown in the correlation matrix, there is a high correlation between c8.06 (Soundness of Banks) and c0.03 (GDP per capita) of 0.7293571

There is also a slight negative correlation between c3.03 (Inflation) and c0.03 (GDP per capita).

This correlation is visualized as follow:

ggplot(df2[df2$Attribute=="Value",],

aes(x=c8.06,

y=c0.03,

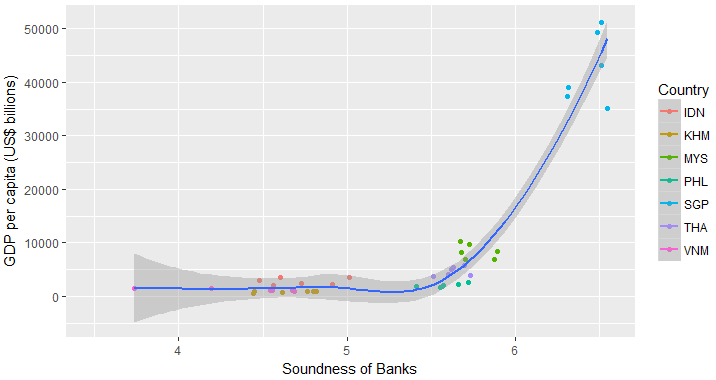
group=1,

color=Country))+

geom\_point()+

geom\_smooth()+

labs(x="Soundness of Banks", y="GDP per capita (US$ billions)")



ggplot(df2[df2$Attribute=="Value",],

aes(x=c8.06,

y=log(c0.03),

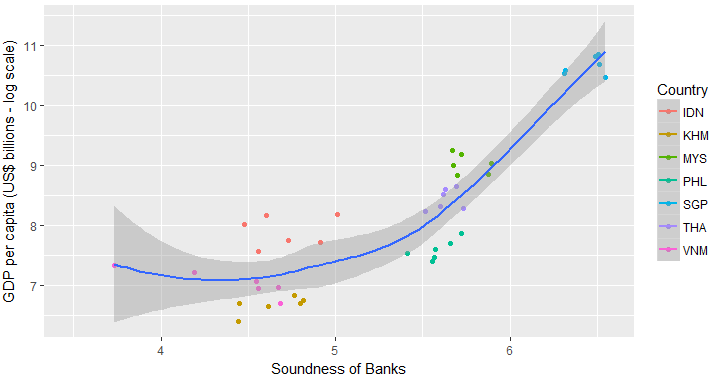
group=1,

color=Country))+

geom\_point()+

geom\_smooth()+

labs(x="Soundness of Banks", y="GDP per capita (US$ billions - log scale)")



ggplot(df2[df2$Attribute=="Value",],

aes(x=c8.06,

y=log(c0.03),

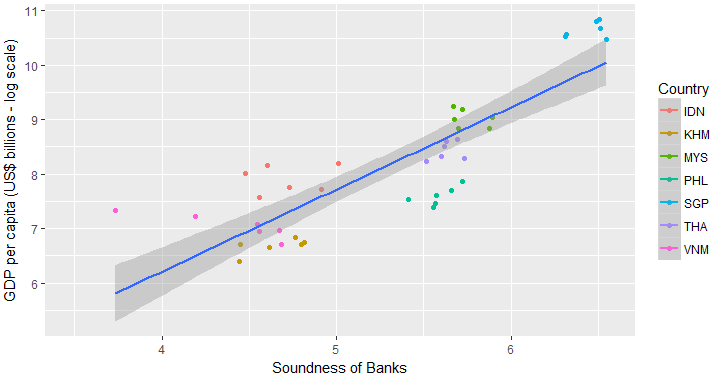
group=1,

color=Country))+

geom\_point()+

geom\_smooth(method="lm")+

labs(x="Soundness of Banks", y="GDP per capita (US$ billions - log scale)")



Soundness of Banks is correlated with the GDP per capita. The higher the soundness of banks, it is likely that the GDP per capita is also higher.