Corruption Index VS Human Development Index Plot

Problem Statement :

To recreate visualization like “The Economist - Corruption and development(http://www.economist.com/node/21541178 (http://www.economist.com/node/21541178)),using HDI(Human Development Index) and CPI(Corruption Perception Index) data .
Sample data(Economist_Assignment_Data.csv) is available in this directory only.

Solution :

This is demonstration of my approach and how have I tried generating a graph mentioned in problem statement.

1. Load data

```
library(data.table)
df_hdi_cpi <- fread('Economist_Assignment_Data.csv',drop=1)
head(df_hdi_cpi)
```

```
##        Country HDI.Rank   HDI CPI            Region
## 1: Afghanistan      172 0.398 1.5      Asia Pacific
## 2:     Albania       70 0.739 3.1 East EU Cemt Asia
## 3:     Algeria       96 0.698 2.9              MENA
## 4:      Angola      148 0.486 2.0               SSA
## 5:   Argentina       45 0.797 3.0          Americas
## 6:     Armenia       86 0.716 2.6 East EU Cemt Asia
```

2. Explore data
   find distinct number of countries, regions, number of countries per regions

```
length(unique(df_hdi_cpi$Region))
```

```
## [1] 6
```

```
length(unique(df_hdi_cpi$Country))
```

```
## [1] 173
```

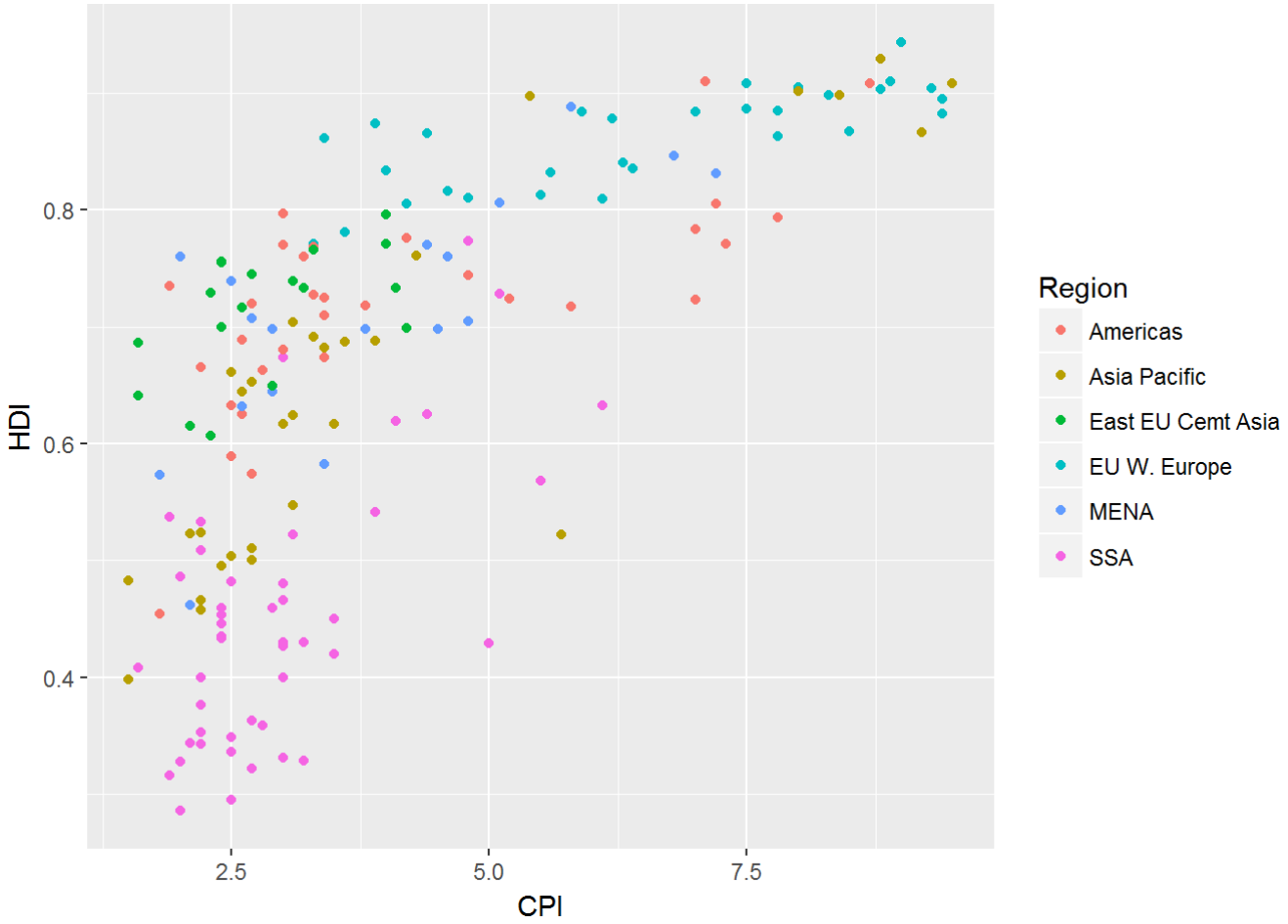
```
library(dplyr)
summarize(group_by(df_hdi_cpi, Region), countriey_count = n()) %>% arrange(countriey_count)
```

```
## # A tibble: 6 × 2
##              Region countriey_count
##               <chr>           <int>
## 1 East EU Cemt Asia              18
## 2              MENA              18
## 3      Asia Pacific              30
## 4      EU W. Europe              30
## 5          Americas              31
## 6               SSA              46
```

There are 2 columns which should be explored - HDI and CPI. This is what is represented in target plot also.

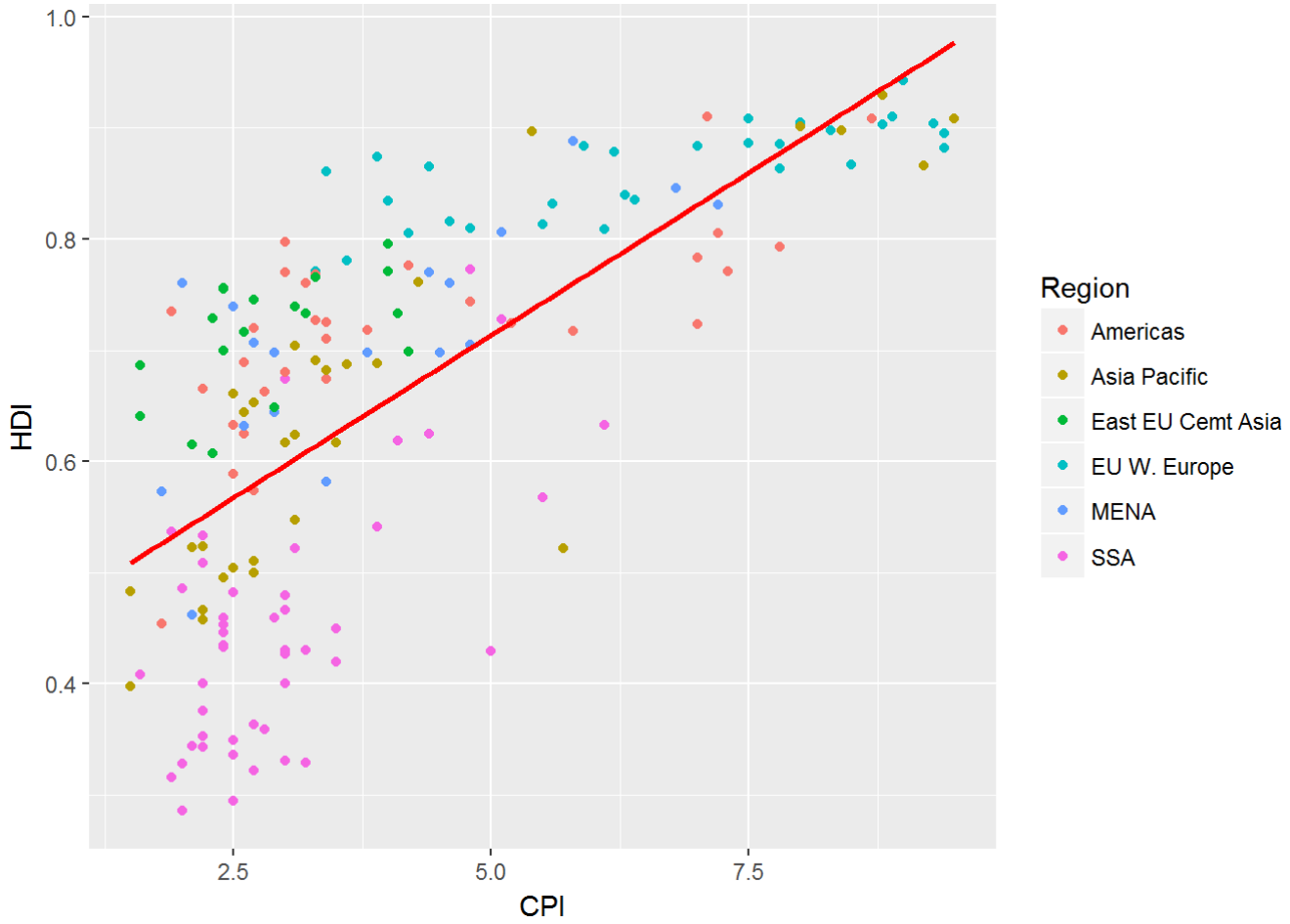
3. Plotting

```
library(ggplot2)
plot_basic <- ggplot(df_hdi_cpi, aes(x = CPI, y = HDI, color = Region)) + geom_point()
print(plot_basic)
```



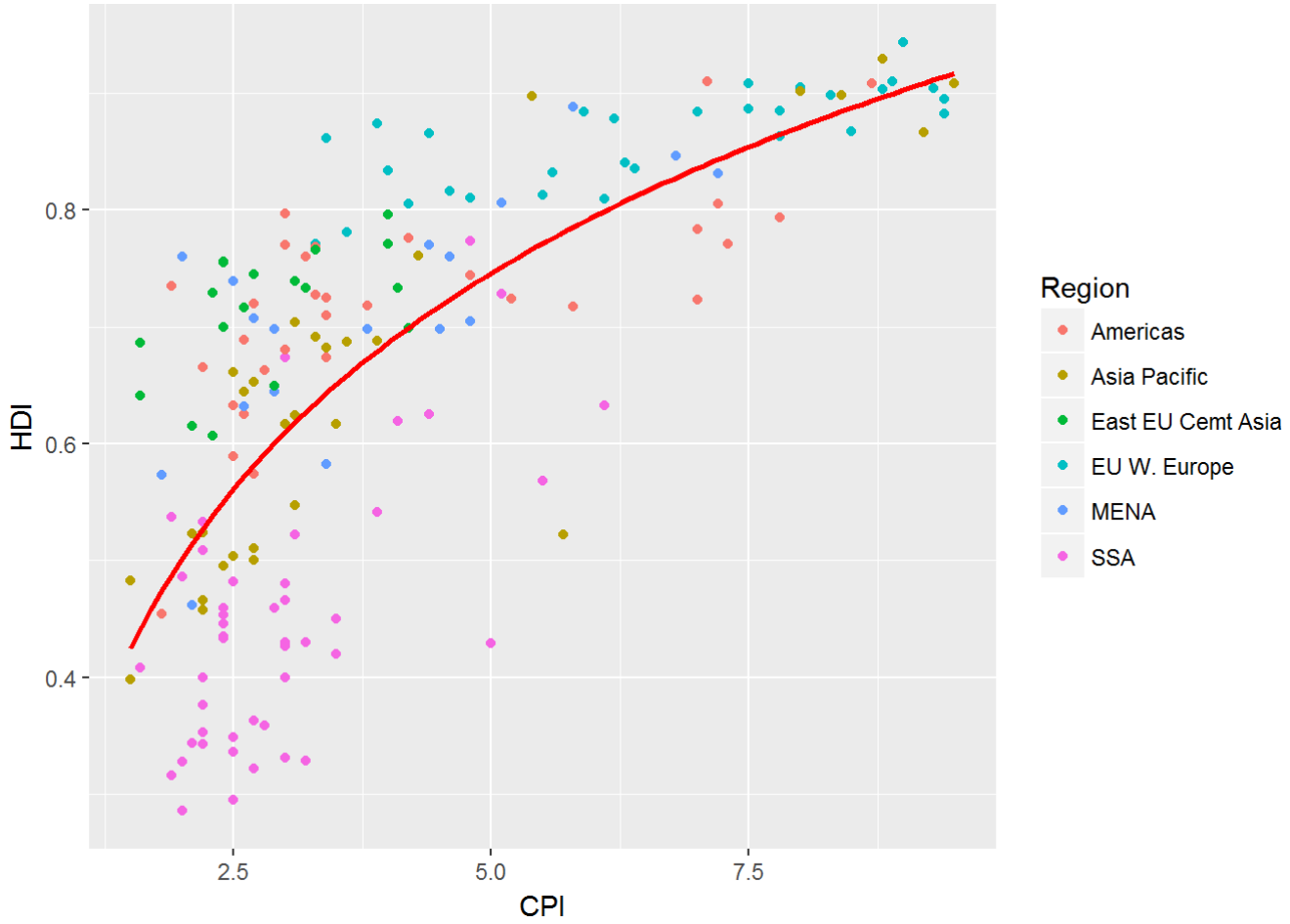
Add trend line - Regression line here - look at final impage that has to be achieved

```
plot_regression <- plot_basic + geom_smooth(aes(group = 1), method = 'lm', formula = y~x, se
= F, color = 'red')
print(plot_regression)
```



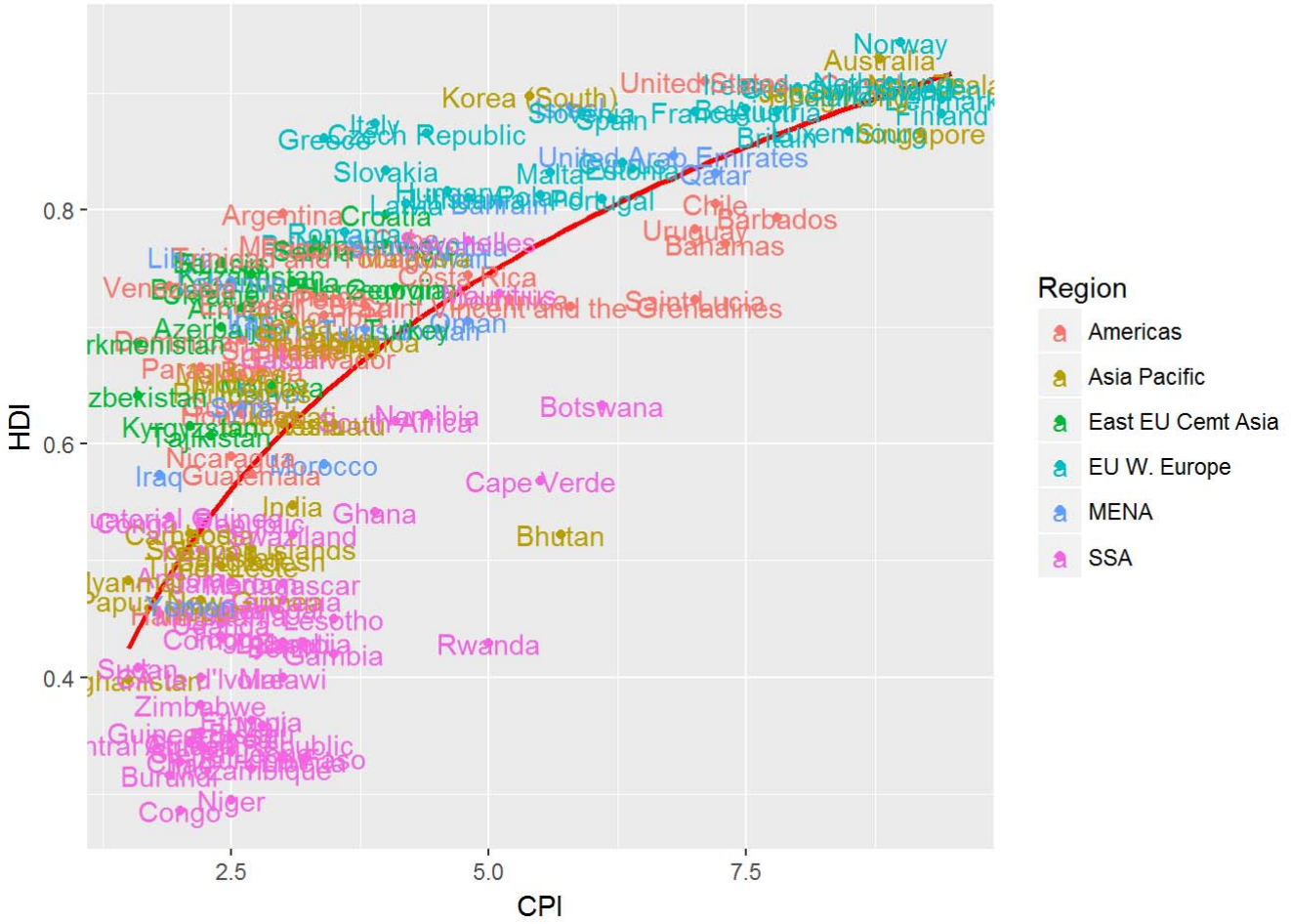
Required output line is not linear, its log

```
plot_regression <- plot_basic + geom_smooth(aes(group = 1), method = 'lm', formula = y ~
log(x), se = F, color = 'red')
print(plot_regression)
```



Adding labels to circles

```
print(plot_regression + geom_text(aes(label=Country)))
```

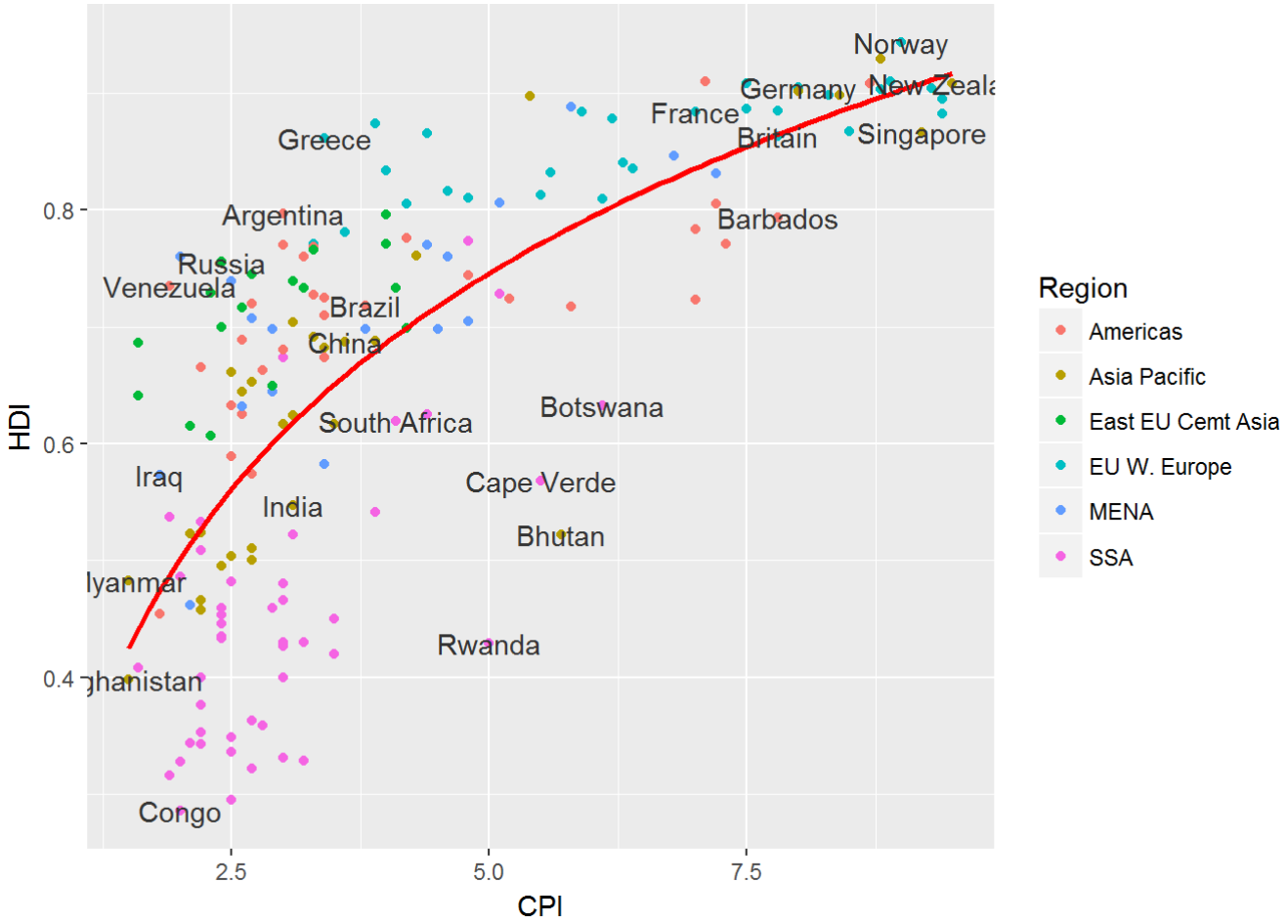


We dont need all countries in plot final image
(Look at the impage and list down names of all countries required)

```
pointsToLabel <- c("Russia", "Venezuela", "Iraq", "Myanmar", "Sudan",
                   "Afghanistan", "Congo", "Greece", "Argentina", "Brazil",
                   "India", "Italy", "China", "South Africa", "Spane",
                   "Botswana", "Cape Verde", "Bhutan", "Rwanda", "France",
                   "United States", "Germany", "Britain", "Barbados", "Norway", "Japan",
                   "New Zealand", "Singapore")
```

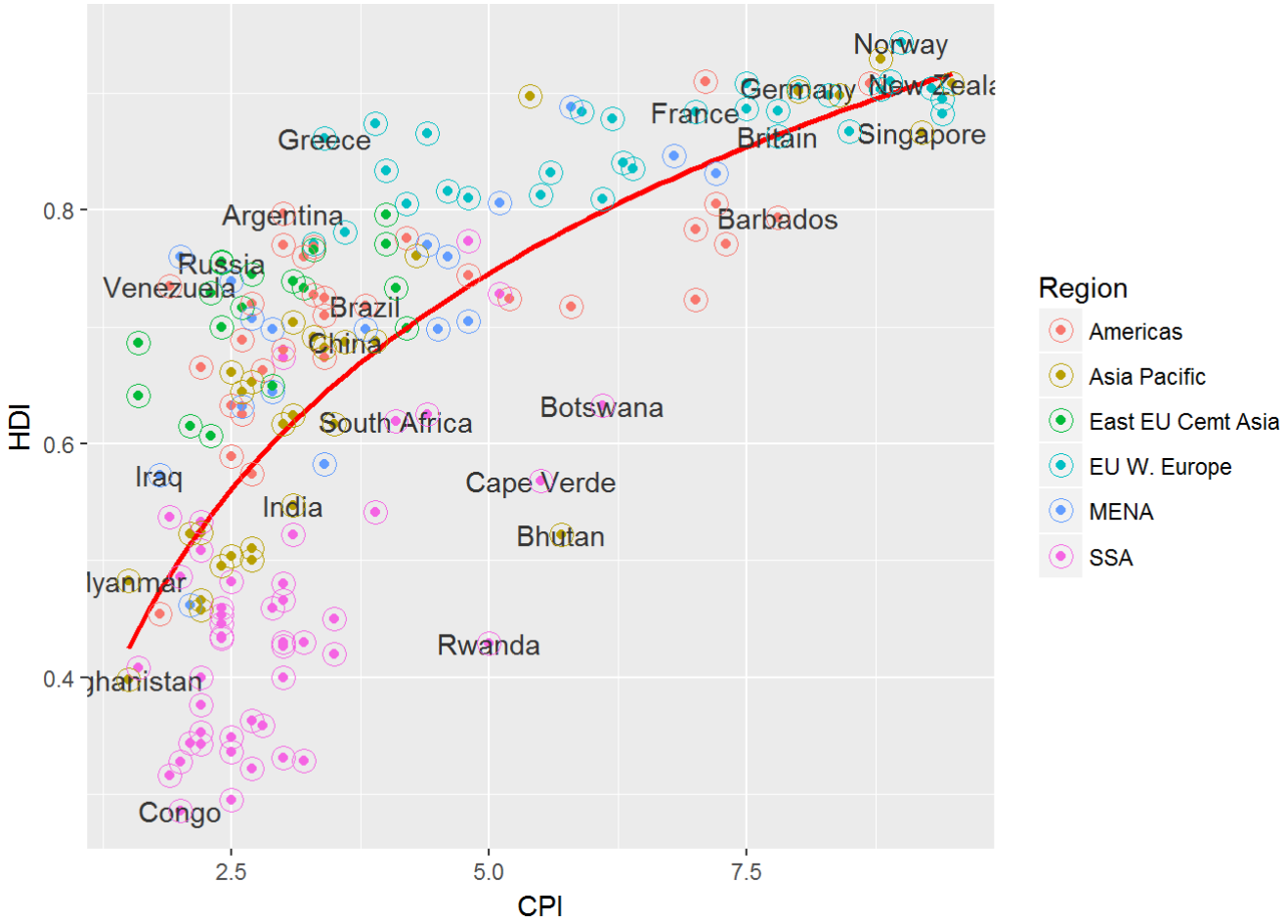
Plot only those countries which are required on final plot

```
plot_required_countries <- plot_regression + geom_text(aes(label = Country), color =
"gray20",
                                                       data = subset(df_hdi_cpi, Country %in%
 pointsToLabel),check_overlap = TRUE)
print(plot_required_countries)
```



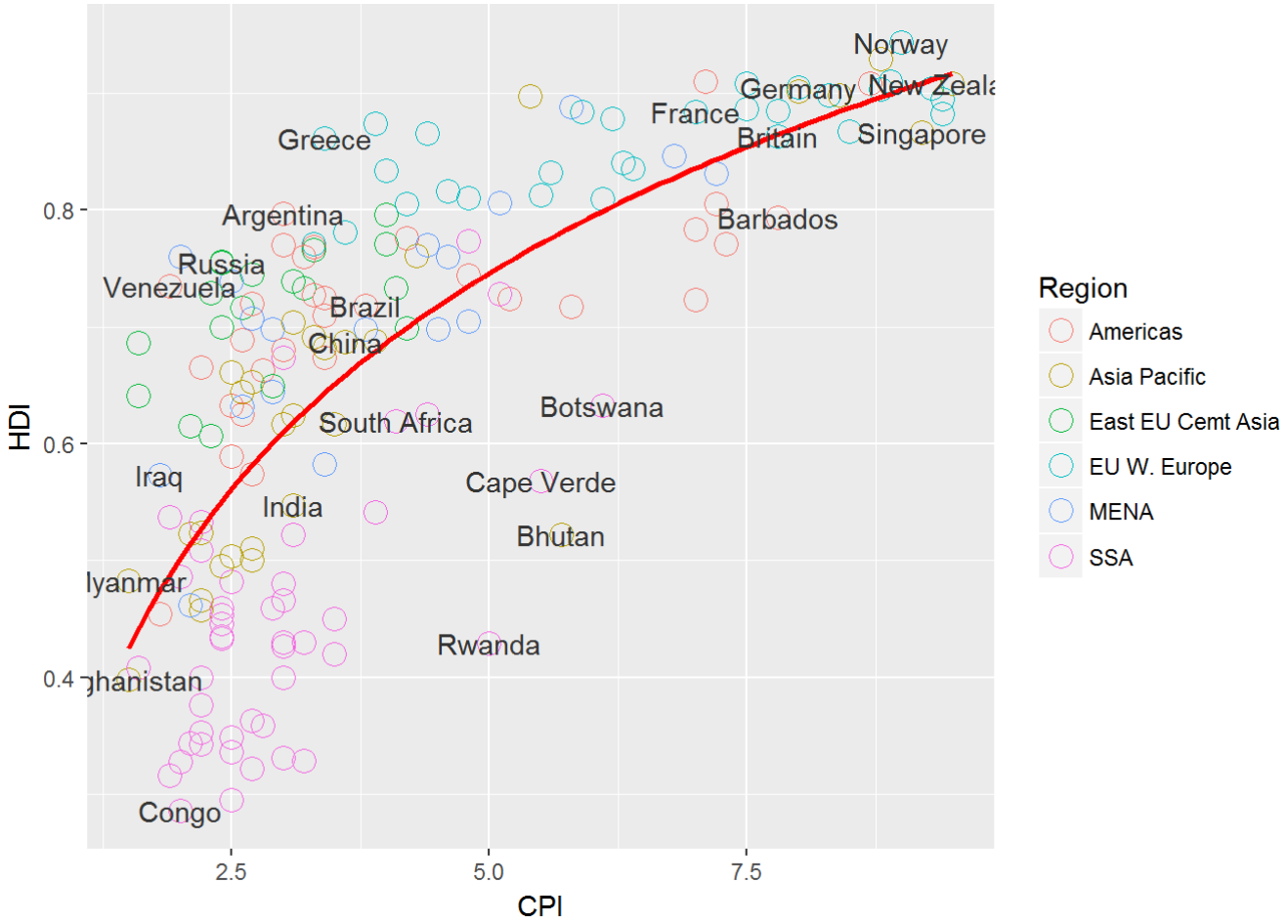
Make circles hollow

```
plot_required_countries <- plot_required_countries + geom_point(size=4,shape=1)
print(plot_required_countries)
```



Not as expected.
apply this to first plot

```
plot_required_countries <- ggplot(df_hdi_cpi, aes(x = CPI, y = HDI, color = Region)) + geom_p
oint(size=4,shape=1) +
        geom_smooth(aes(group = 1), method = 'lm', formula = y ~ log(x), se = F, color = 're
d') +
        geom_text(aes(label = Country), color = "gray20", data = subset(df_hdi_cpi, Country %
in% pointsToLabel),check_overlap = TRUE)
print(plot_required_countries)
```

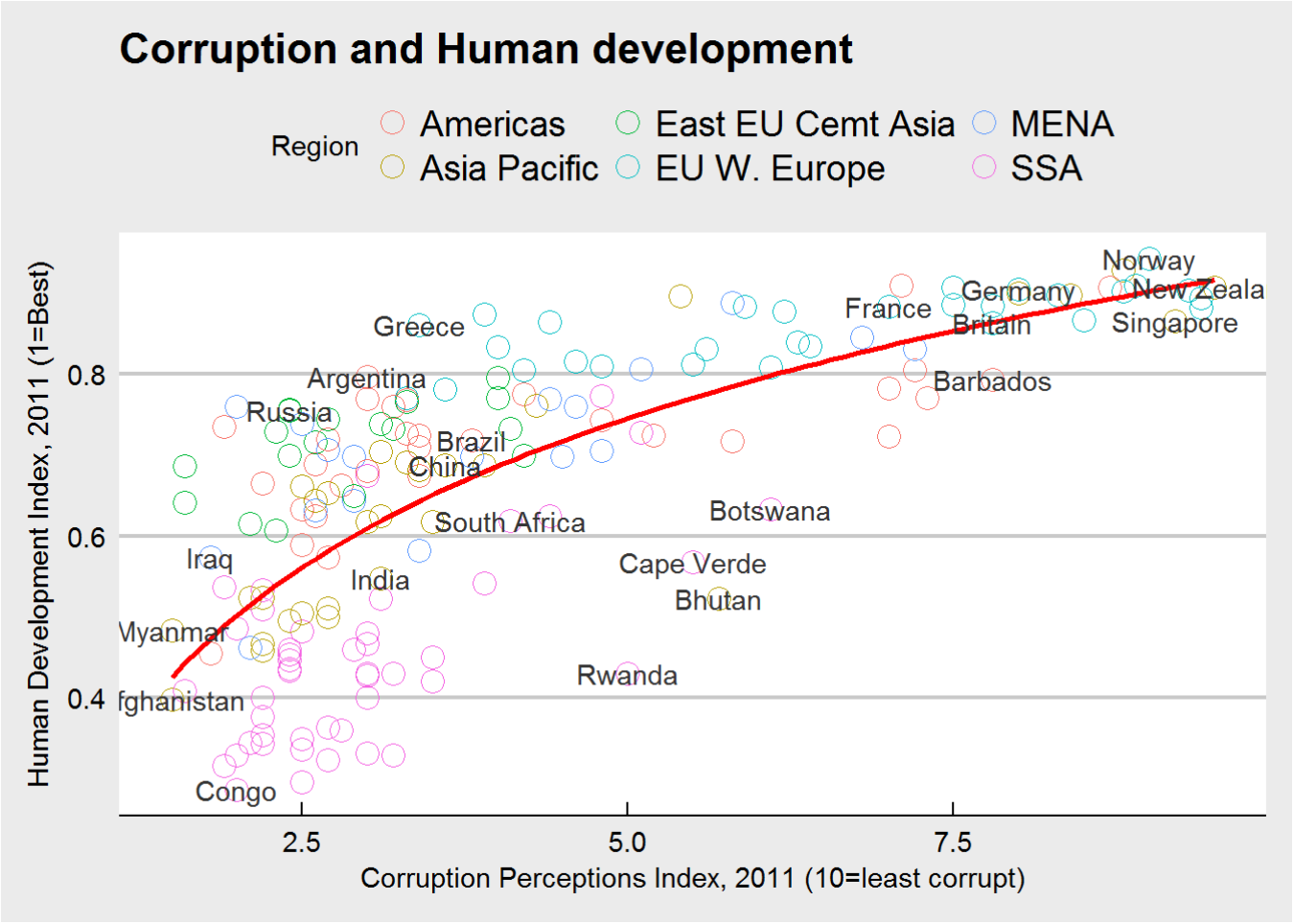


Adding headings etc

```
plot_x_y_axis <- plot_required_countries + scale_x_continuous(name = "Corruption Perceptions
 Index, 2011 (10=least corrupt)" ) + scale_y_continuous(name = "Human Development Index, 2011
 (1=Best)")
```

Adding upper heading + theme

```
library(ggthemes)
plot_hdi_cpi_final <- plot_x_y_axis + ggtitle("Corruption and Human development") + theme_eco
nomist_white()
print(plot_hdi_cpi_final)
```


Corruption and Human development
Region
Americas
East EU Cemt Asia
MENA
Asia Pacific
EU W. Europe
SSA
Norway
New Zeala
Germany
France
Britain
Singapore
Greece
Argentina
Barbados
Russia
Brazil
China
South Africa
Botswana
Iraq
India
Cape Verde
Bhutan
Myanmar
Rwanda
fghanistan
Congo
0.8
0.6
0.4
2.5
5.0
7.5
Human Development Index, 2011 (1=Best)
Corruption Perceptions Index, 2011 (10=least corrupt)