Economist

Hanh Nguyen

The dataset (Economist.csv) consists of countries scored on how corrupt their public sectors are seen to be (Corruption Perceptions Index - CPI) and on achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living (Human Development Index - HDI).
Note: CPI scale goes from 0 (highly corrupt) to 10 (very clean).

```
library(ggplot2)
library(ggrepel)
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
```

```
library(reshape2)
dat = read.csv(file="/Users/user/GitHub/data-vis/dataSets/EconomistData.csv",header=TRUE)
```

```
head(dat)
```

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

```
str(dat)
```

```
## 'data.frame':    173 obs. of  6 variables:
##  $ X       : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country : Factor w/ 173 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ HDI.Rank: int  172 70 96 148 45 86 2 19 91 53 ...
##  $ HDI     : num  0.398 0.739 0.698 0.486 0.797 0.716 0.929 0.885 0.7 0.771 ...
##  $ CPI     : num  1.5 3.1 2.9 2 3 2.6 8.8 7.8 2.4 7.3 ...
##  $ Region  : Factor w/ 6 levels "Americas","Asia Pacific",..: 2 3 5 6 1 3 2 4 3 1 ...
```

```
summary(dat)
```

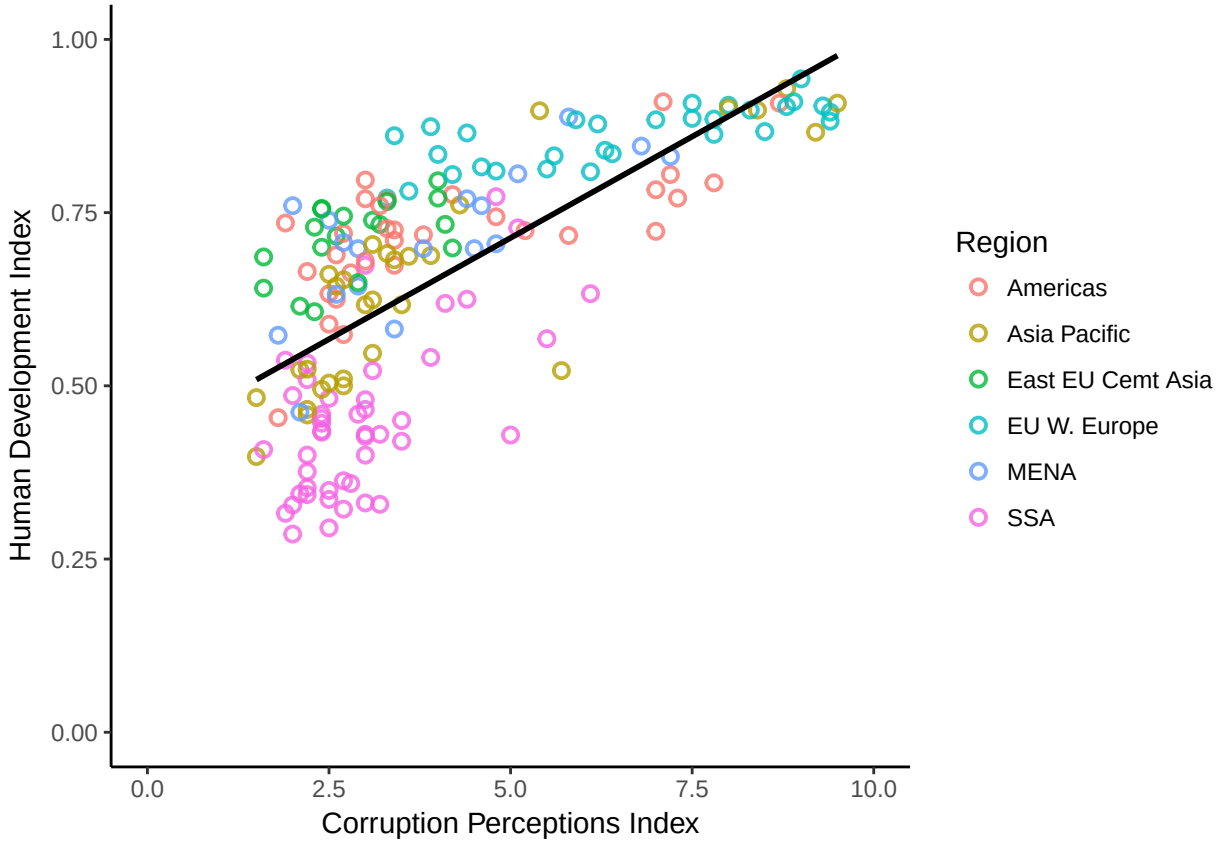
```
##        X              Country       HDI.Rank          HDI
##  Min.   :  1   Afghanistan:  1   Min.   :  1.00   Min.   :0.2860
##  1st Qu.: 44   Albania    :  1   1st Qu.: 47.00   1st Qu.:0.5090
##  Median : 87   Algeria    :  1   Median : 96.00   Median :0.6980
##  Mean   : 87   Angola     :  1   Mean   : 95.28   Mean   :0.6581
##  3rd Qu.:130   Argentina  :  1   3rd Qu.:143.00   3rd Qu.:0.7930
##  Max.   :173   Armenia    :  1   Max.   :187.00   Max.   :0.9430
```

```
##                 (Other)    :167
##       CPI                        Region
##  Min.   :1.500   Americas          :31
##  1st Qu.:2.500   Asia Pacific      :30
##  Median :3.200   East EU Cemt Asia:18
##  Mean   :4.052   EU W. Europe      :30
##  3rd Qu.:5.100   MENA              :18
##  Max.   :9.500   SSA               :46
##
```

1. Plotting HDI and CPI

A scatterplot can show how countries are measured in terms of corruption and human development.

```
dat %>%
  ggplot(aes(x = CPI, y = HDI)) +
  geom_point(aes(col=Region),shape=1,stroke=1,size=2,alpha=.8) +
  geom_smooth(method="lm",se=FALSE,col="black") +
  labs(x="Corruption Perceptions Index",y="Human Development Index") +
  theme_classic() +
  xlim(0,10) + ylim(0,1)
```



Note: geom_smooth() is used to add a smooth line.
The plot indicates a positive correlation between HDI and CPI. In fact, their correlation is 0.7 which is fairly high.

```
cor.test(dat$CPI,dat$HDI)
```

```
##
##  Pearson's product-moment correlation
##
## data:  dat$CPI and dat$HDI
## t = 12.994, df = 171, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6209764 0.7727980
## sample estimates:
##       cor
## 0.7048705
```

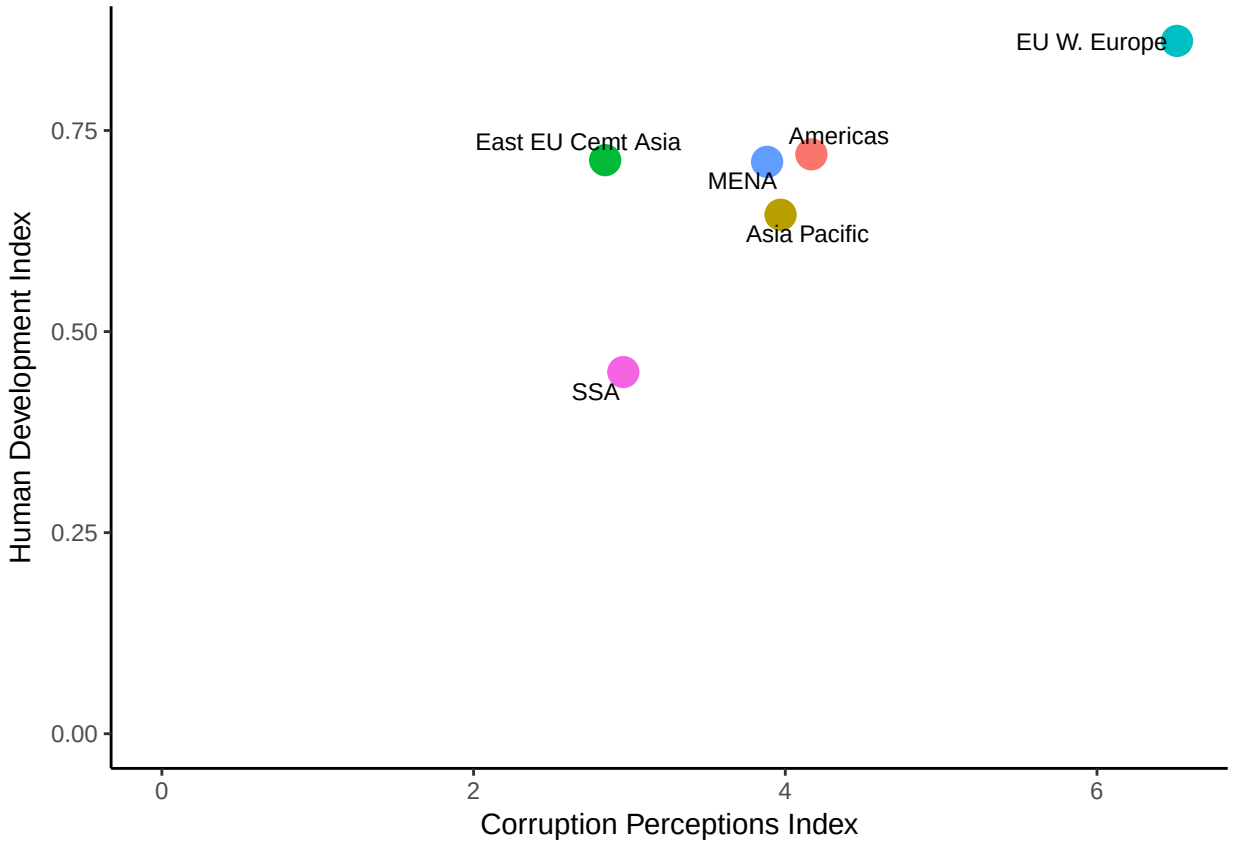
To see how each region performs, we first aggregate (group by) the data by region

```
reg_dat= dat %>%
  group_by(Region) %>%
  summarize(avgCPI=mean(CPI,na.rm=T),avgHDI=mean(HDI,na.rm=T))
reg_dat
```

```
## # A tibble: 6 × 3
##              Region   avgCPI    avgHDI
##              <fctr>    <dbl>     <dbl>
## 1          Americas 4.167742 0.7203226
## 2      Asia Pacific 3.970000 0.6452667
## 3 East EU Cemt Asia 2.844444 0.7131111
## 4      EU W. Europe 6.513333 0.8613667
## 5              MENA 3.883333 0.7110556
## 6               SSA 2.960870 0.4496739
```

The code is similar to the code plotting countries

```
reg_dat %>%
  ggplot(aes(x = avgCPI, y = avgHDI)) +
  geom_point(aes(col=Region),size=5) +
  labs(x="Corruption Perceptions Index",y="Human Development Index") +
  theme_classic() +
  geom_text_repel(aes(avgCPI, avgHDI, label = Region),size=3) +
  expand_limits(x = 0, y = 0) +
  guides(col=F)
```



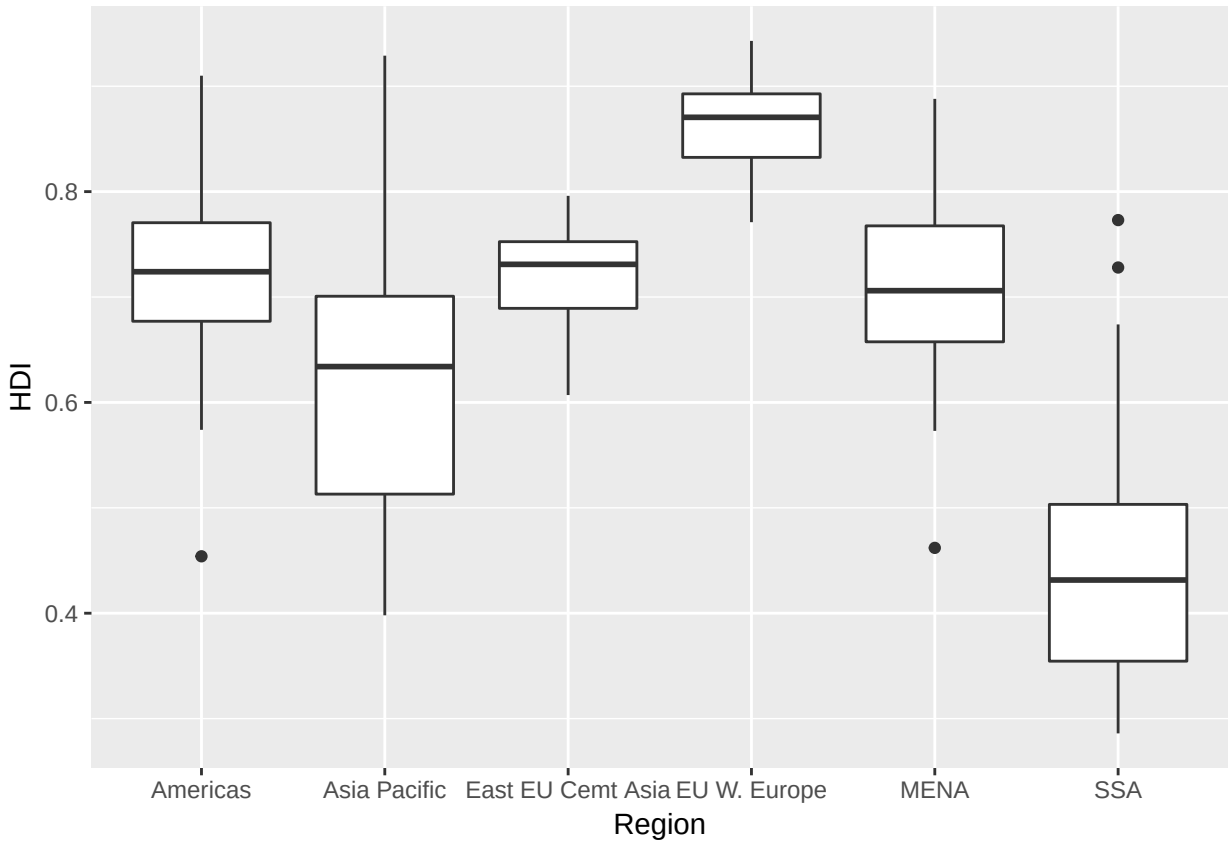
Countries are classified into 6 regions: Western Europe, Americas, Middle East and North Africa (MENA), Asia Pacific, Eastern Europe, and Sub-Saharan Africa (SSA).

Western Europe has the highest HDI and CPI. Sub-Saharan Africa, in contrast, has the lowest HDI and CPI. Americas, Asia Pacific and Middle East and North Africa area are comparable in two indexes.

2. Analyzing HDI and CPI

A box plot can describe more information, including distribution, average and variability.

```
dat %>%
  ggplot(aes(x = Region, y = HDI)) + geom_boxplot()
```



Western EU and Eastern EU have comparatively short boxplots, meaning that countries share similar standards of human development within each region. Americas and Eastern EU have an equal median but different distributions. Asia Pacific has a tall boxplot, indicating that there's a big gap in HDI among countries.

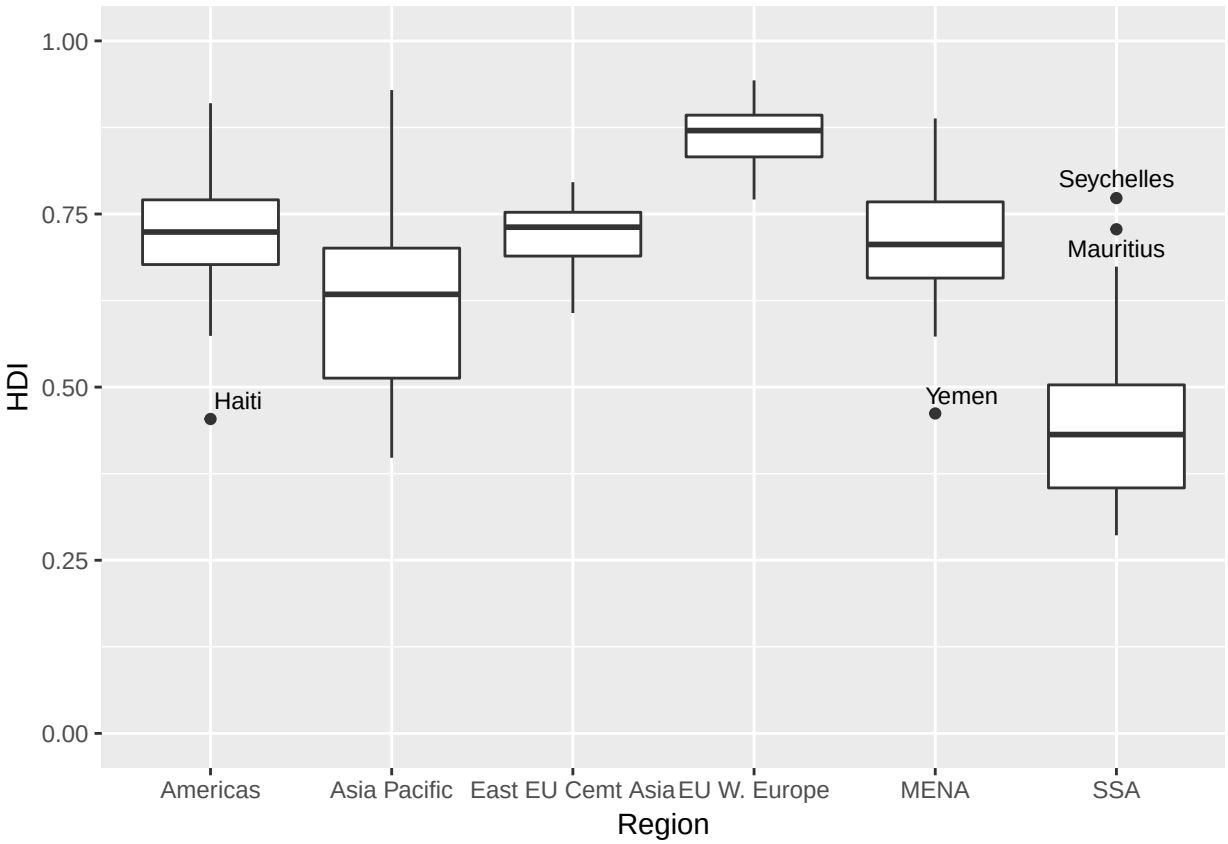
Also, there're a few outliers and we would like to know which countries are they.

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}

dat2 = dat %>%
  group_by(Region) %>%
  mutate(is_outlier = ifelse(is_outlier(HDI), HDI, as.numeric(NA)))

dat2$Country[which(is.na(dat2$is_outlier))] <- as.numeric(NA)

dat2 %>%
  ggplot(aes(x = Region, y = HDI)) +
  geom_boxplot() +
  geom_text_repel(aes(label = Country), size=3,na.rm = TRUE) +
  ylim(0,1)
```



From the plot, Haiti has an extremely low HDI compared to other countries in the America continent. Similiar to Yemen in Middle East and North Africa area (MENA). In Sub-Saharan Africa, which consists of all African countries that are fully or partially located south of the Sahara, Seychelles and Mauritius outperform others in HDI.

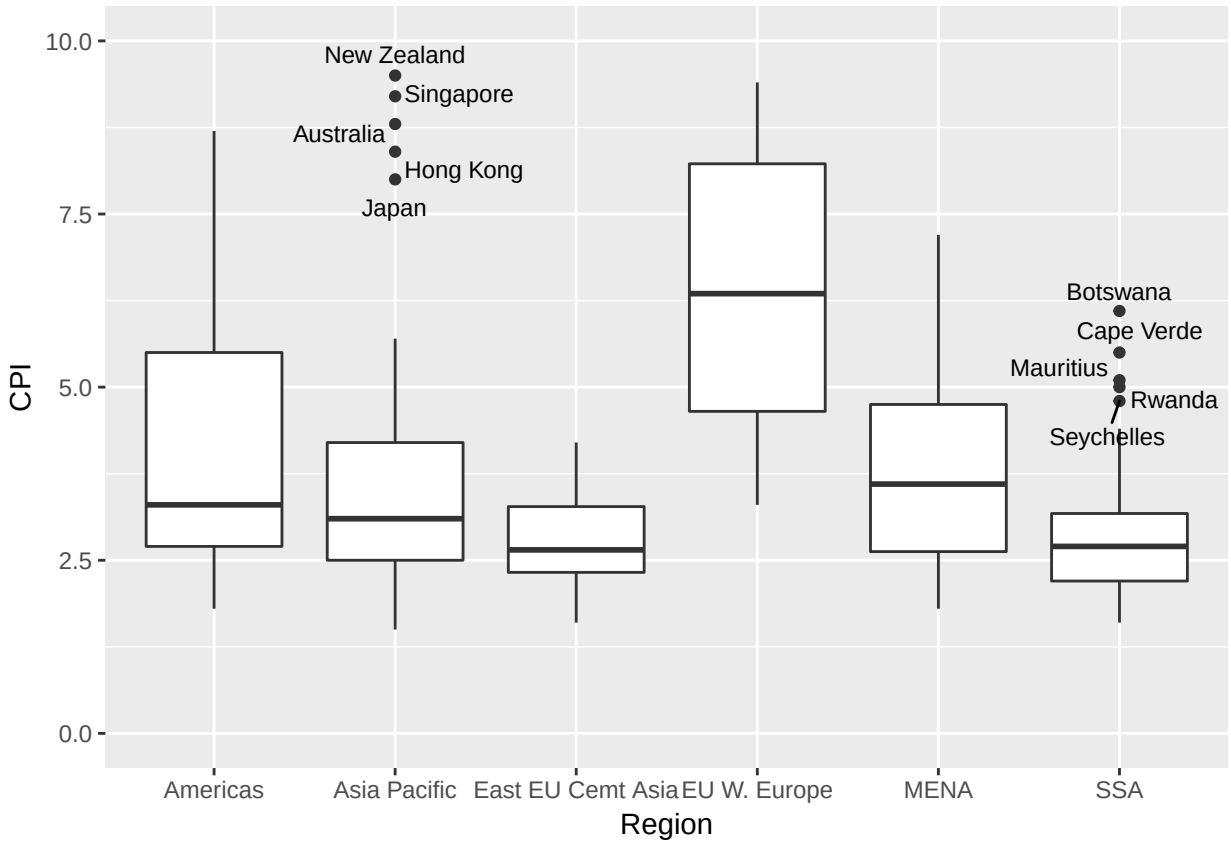
A similar scatterplot for CPI

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}

dat2 = dat %>%
  group_by(Region) %>%
  mutate(is_outlier = ifelse(is_outlier(CPI), CPI, as.numeric(NA)))

dat2$Country[which(is.na(dat2$is_outlier))] <- as.numeric(NA)

dat2 %>%
  ggplot(aes(x = Region, y = CPI)) +
  geom_boxplot() +
  geom_text_repel(aes(label = Country), size=3,na.rm = TRUE) +
  ylim(0,10)
```



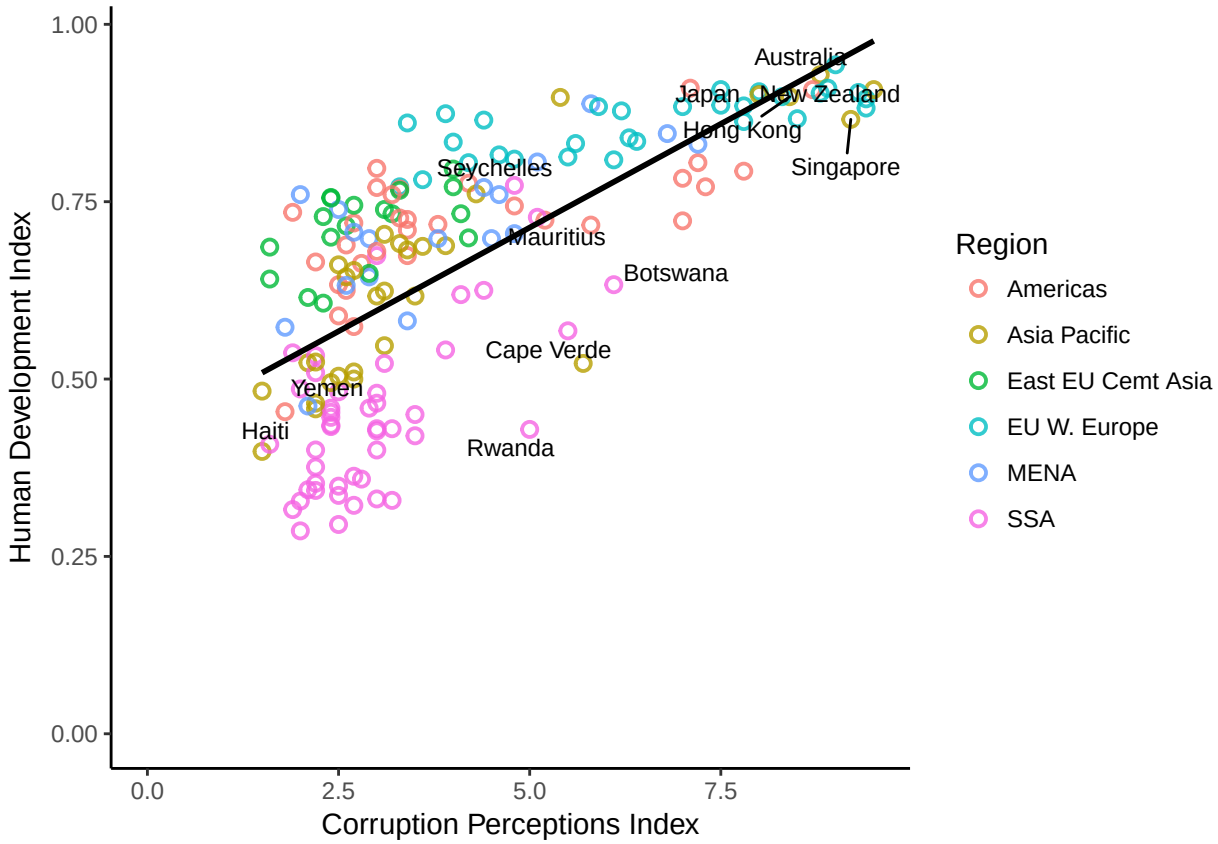
The 4 sections of the box plot of Americas are uneven in size. This shows that countries have similar CPI at the lower quartile groups, but in upper quartiles countries have varied CPI. We also see a few "clean" countries in Asia Pacific and South Africa that stand out from their geographical peers.

Now we want to mark those outliers in the country scatterplot in the beginning.

```
is_outlier <- function(x) {
  return(x < quantile(x, 0.25) - 1.5 * IQR(x) | x > quantile(x, 0.75) + 1.5 * IQR(x))
}
dat1 = dat %>%
  group_by(Region) %>%
  mutate(is_outlier1 = ifelse(is_outlier(CPI), CPI, as.numeric(NA))) %>%
  mutate(is_outlier2 = ifelse(is_outlier(HDI), HDI, as.numeric(NA)))

dat1$Country[which(is.na(dat1$is_outlier1) & is.na(dat1$is_outlier2))] <- as.numeric(NA)
```

```
dat1 %>%
  ggplot(aes(x = CPI, y = HDI)) +
  geom_point(aes(col=Region),shape=1,stroke=1,size=2,alpha=.8) +
  geom_smooth(method="lm",se=FALSE,col="black") +
  labs(x="Corruption Perceptions Index",y="Human Development Index") +
  theme_classic() +
  geom_text_repel(aes(label = Country), size=3,na.rm = TRUE) +
  expand_limits(x = 0, y = 0)
```



Sources:
http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html#org93999d8
https://www.transparency.org/research/cpi/overview
http://hdr.undp.org/en/content/human-development-index-hdi
https://stackoverflow.com/questions/33524669/labeling-outliers-of-boxplots-in-r
https://www.r-bloggers.com/from-continuous-to-categorical/
https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots