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 100 (very clean).

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

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

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

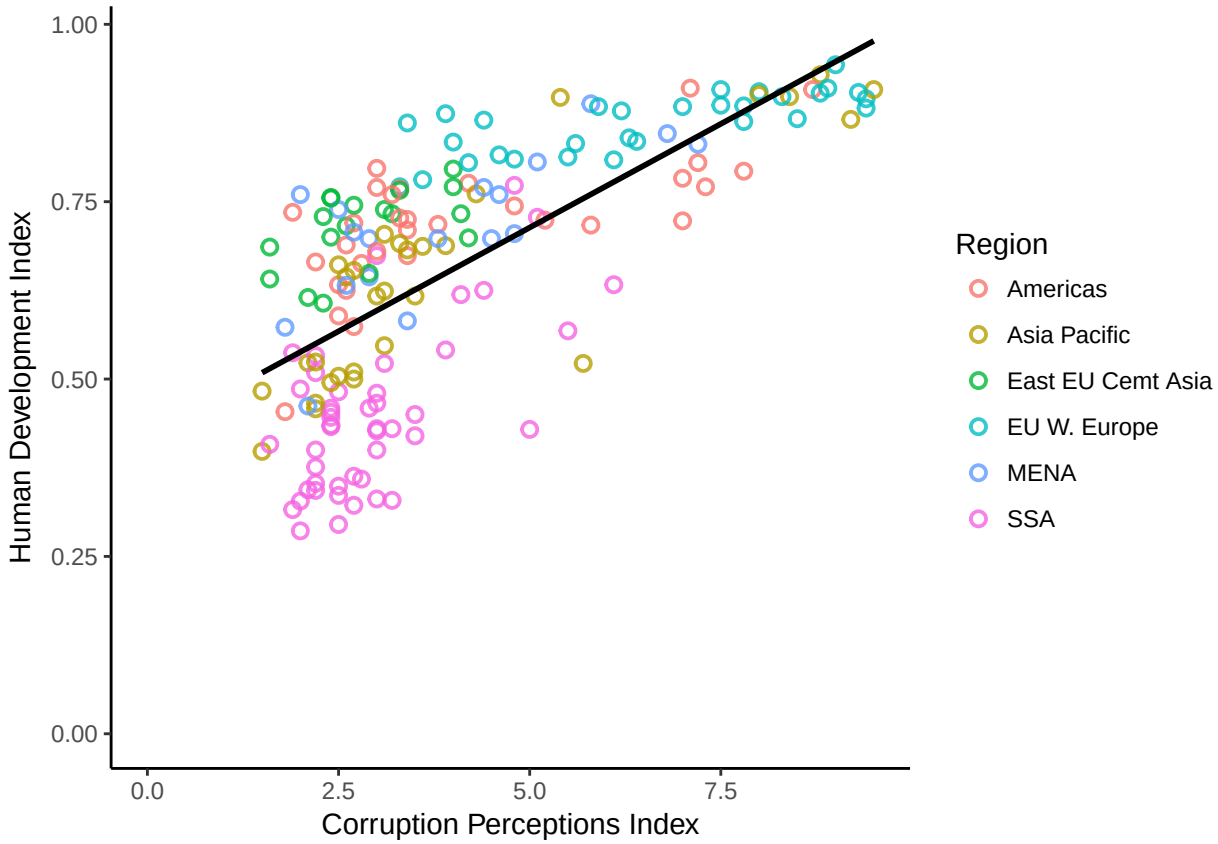
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

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() +
  expand_limits(x = 0, y = 0)
```



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
```

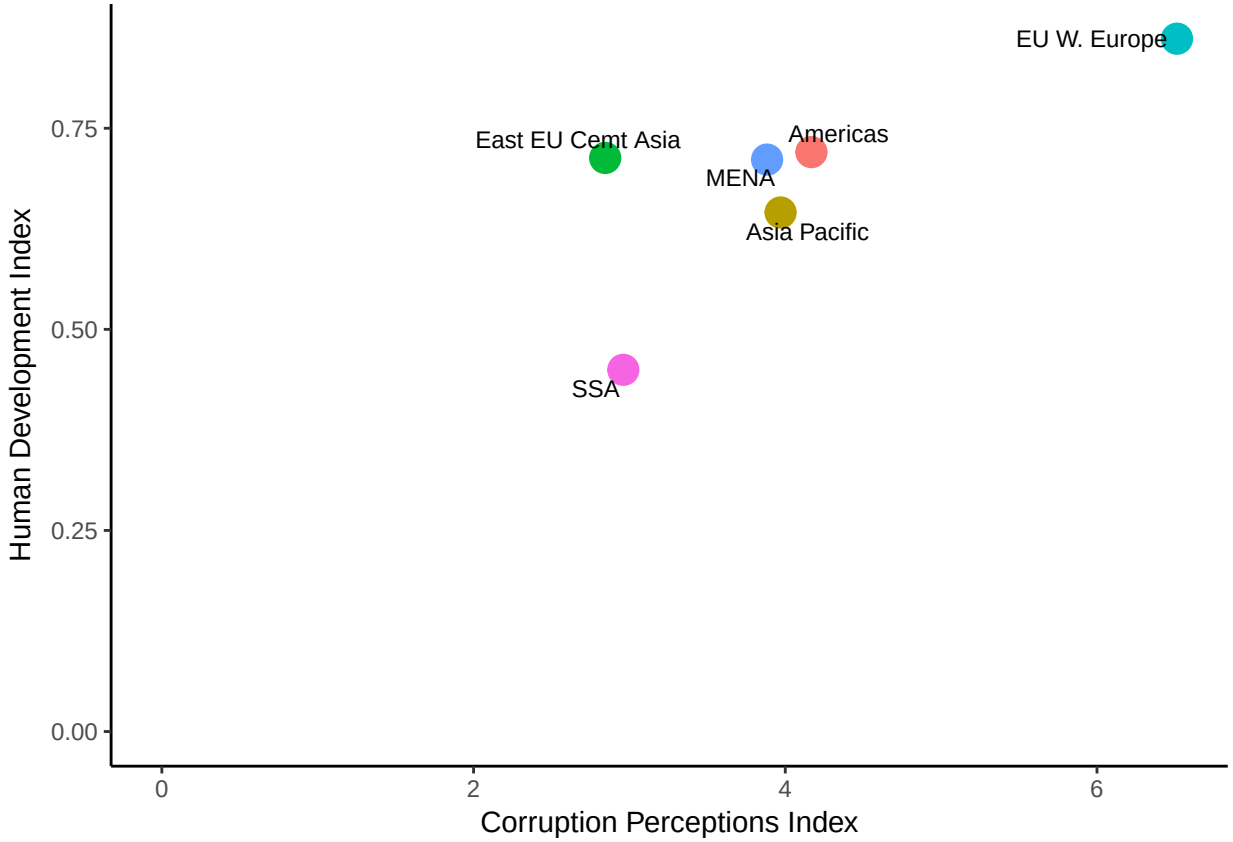
To see how each region performs, we first aggregate (group by) the data by region

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

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)
```

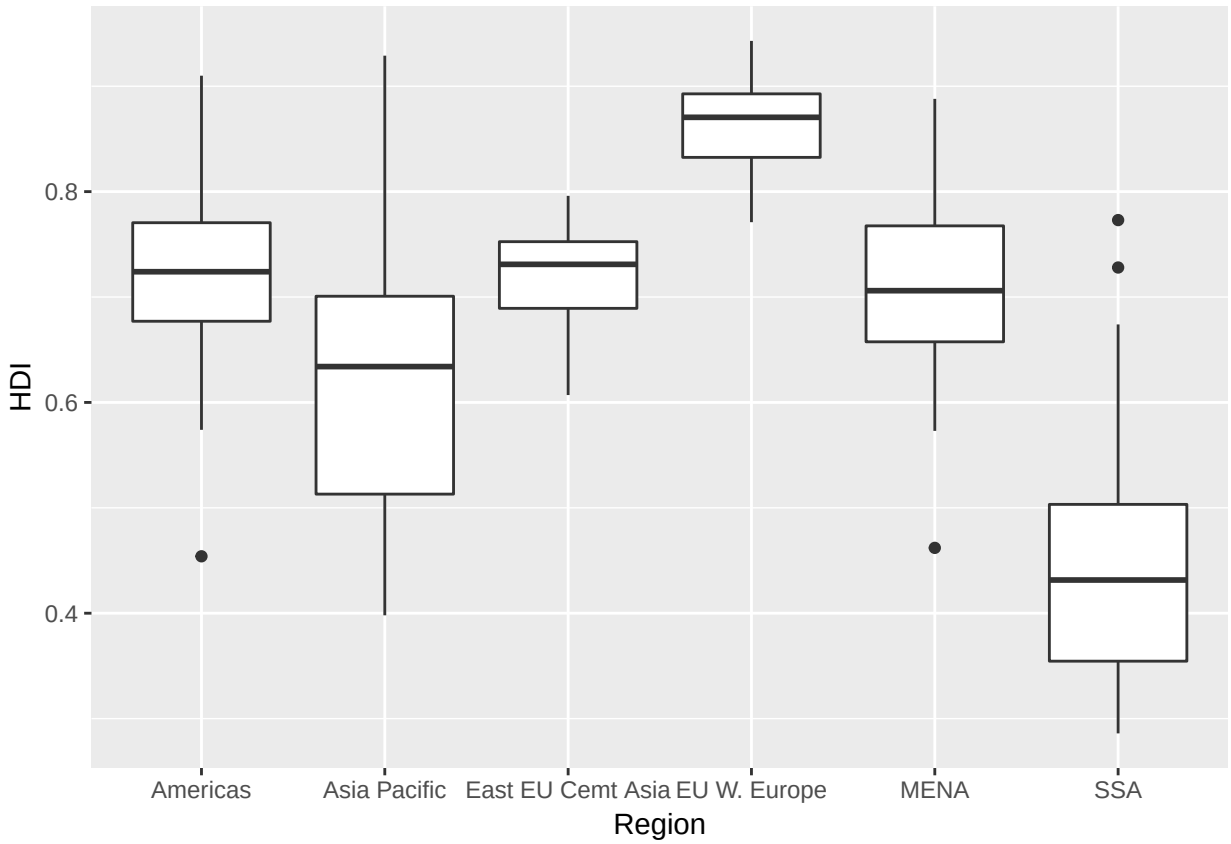


Western Europe has the highest HDI and CPI. Sub-Saharan Africa, in contrast, has the lowest HDI and CPI. Americas, Asia Pacific and Middle East and North Africa area are comparable in two indexes.

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()
```



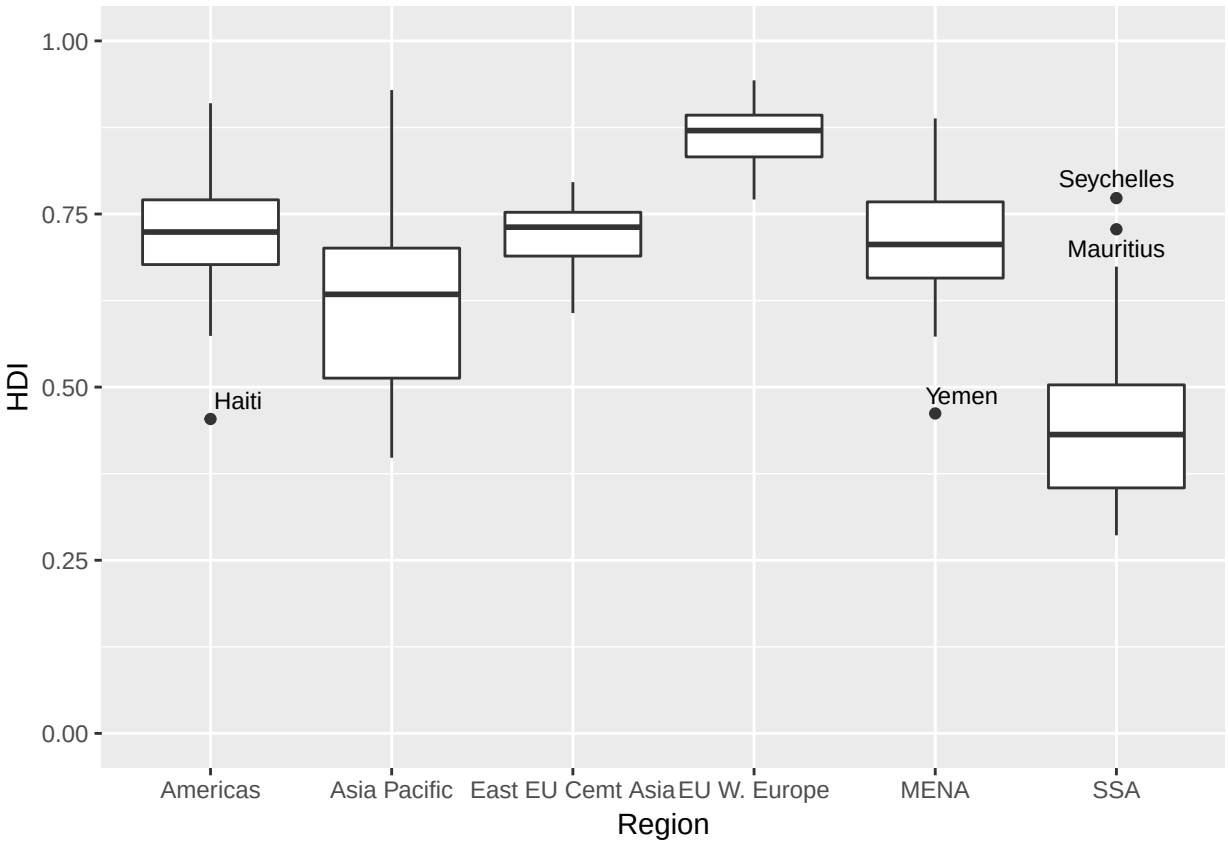
There're a few outliers and we would like to know which countries are they.

```
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.

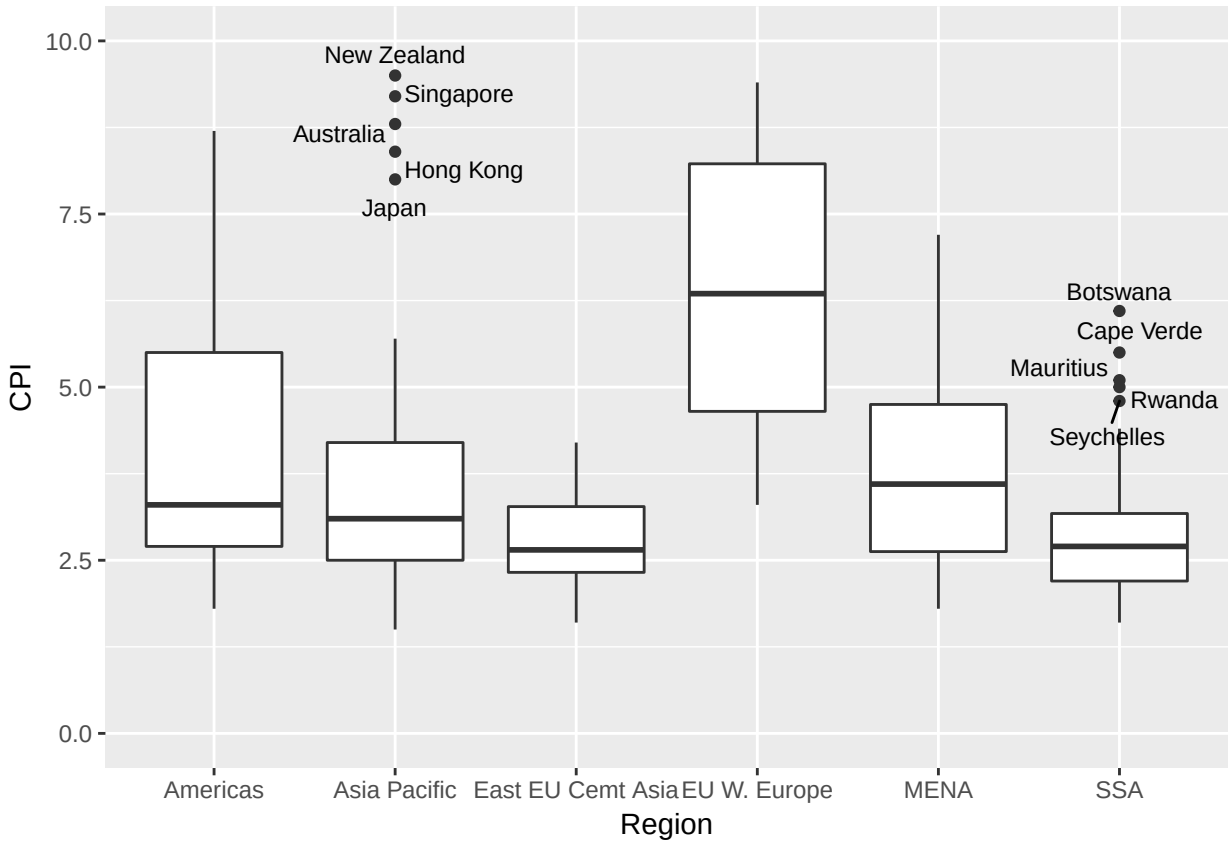
A similar scatterplot for CPI

```
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)
```

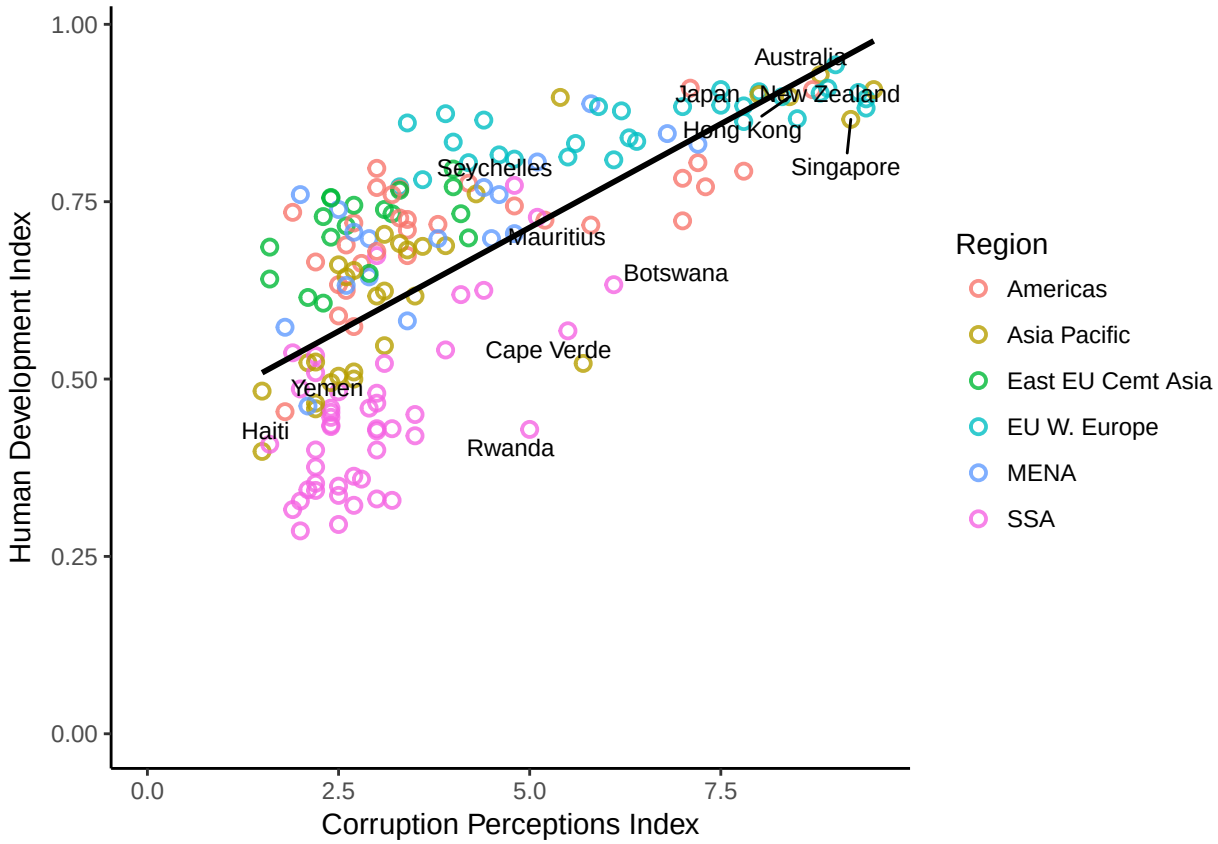


Now we want to mark those outliers in the country scatterplot in the beginning.

```
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/