

# Economist

*Hanh Nguyen*

*5/21/2017*

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).

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

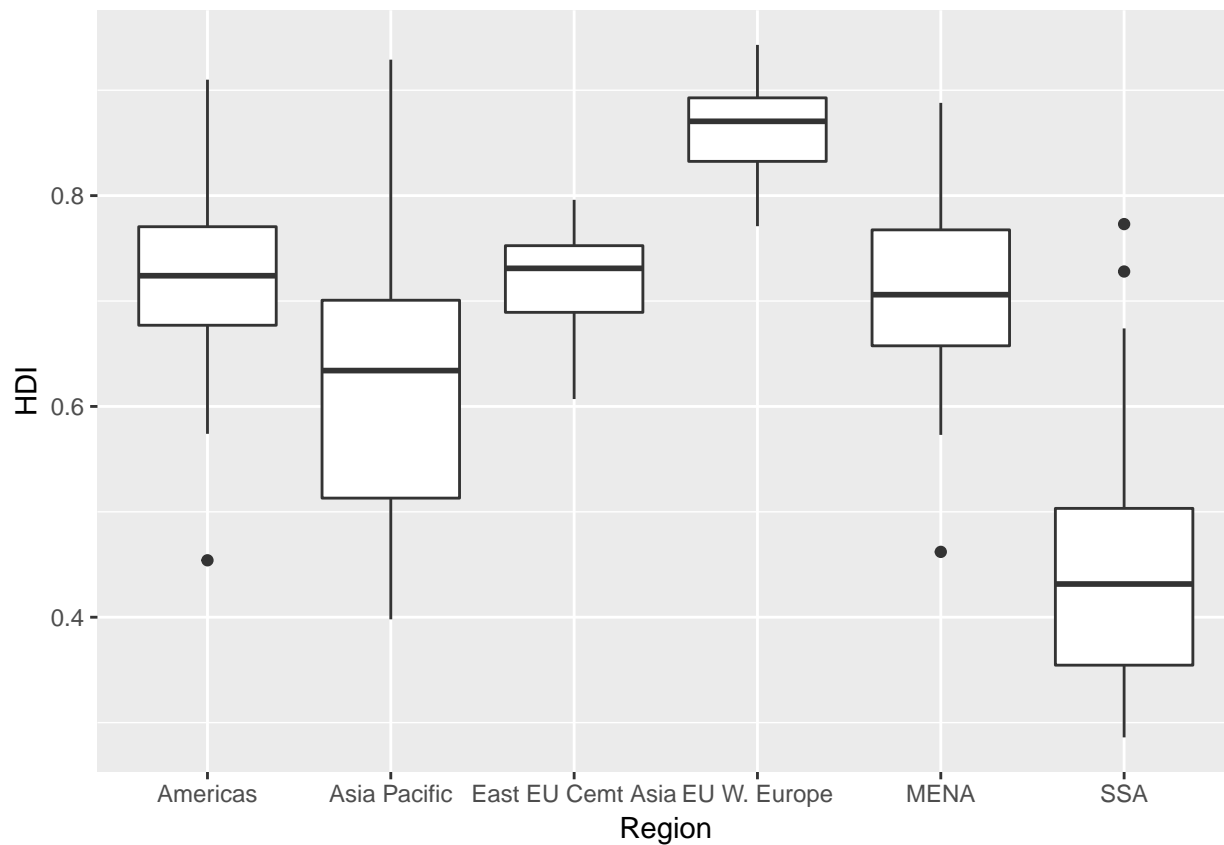
##
## Attaching package: 'dplyr'
## The following objects are masked from 'package:stats':
##
##   filter, lag
## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(reshape2)
dat = read.csv(file="/Users/user/GitHub/data-vis/dataSets/EconomistData.csv",header=TRUE)
dat$CPI.Rank = rank(dat$CPI,ties.method="first")
```

## 1. Visualization on national level

### Human Development summary

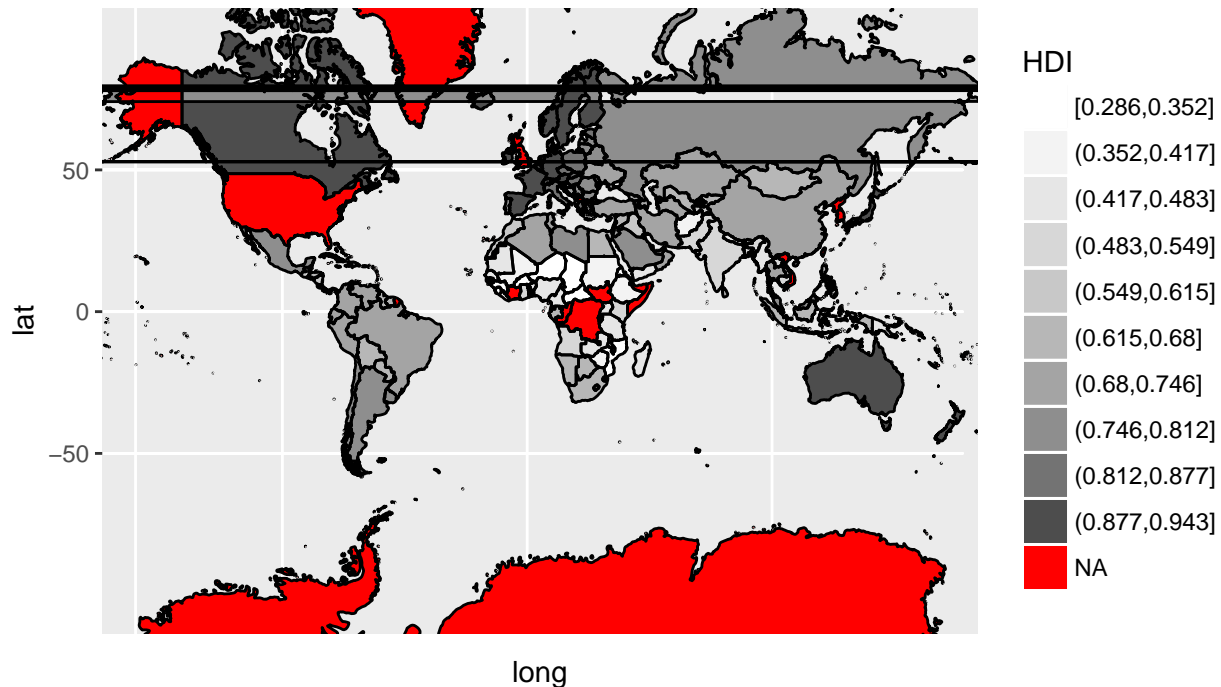
```
dat %>%
  ggplot(aes(x = Region, y = HDI)) + geom_boxplot()
```



## Human Development on the world map

```
dfworldmap = map_data("world")
map_data = merge(dfworldmap, dat, by.x="region", by.y="Country", all=TRUE)
map_data = transform(map_data, HDI = cut_interval(HDI, 10))
map_data = map_data[order(map_data$order),]
map_data %>%
  ggplot(aes(x=long, y=lat, group=group)) +
  geom_polygon(aes(fill = HDI)) +
  scale_fill_grey(start=1, end=.3) +
  geom_path(colour='black') +
  coord_map() +
  ggtitle("Human Development Index Ranking of the world")
```

## Human Development Index Ranking of the world



*cut\_interval* makes n groups with equal range; *cut\_number* makes n groups with (approximately) equal numbers of observations; *cut\_width* makes groups of width.

Some countries are named differently from *dat* and *dfworldmap* so renaming them is necessary.

Countries in *dat* that do not have any match in *dfworldmap*: `df1 = map_data[is.na(map_data$long),]`

Countries in *dfworldmap* that do not have any match in *dfworldmap*: `df2 = map_data[is.na(map_data$CPI),]` and `sort(unique(df2$region))`

```
dat$Country[dat$Country=="Korea (South)"] = "South Korea"
```

```
## Warning in `[<-.factor`(`*tmp*`, dat$Country == "Korea (South)", value =  
## structure(c(1L, : invalid factor level, NA generated
```

```
dat$Country[dat$Country=="United States"] = "USA"
```

```
## Warning in `[<-.factor`(`*tmp*`, dat$Country == "United States", value =  
## structure(c(1L, : invalid factor level, NA generated
```

```
dat$Country[dat$Country=="Britain"] = "UK"
```

```
## Warning in `[<-.factor`(`*tmp*`, dat$Country == "Britain", value =  
## structure(c(1L, : invalid factor level, NA generated
```

```
dat$Country[dat$Country=="Trinidad and Tobago"] = "Trinidad"
```

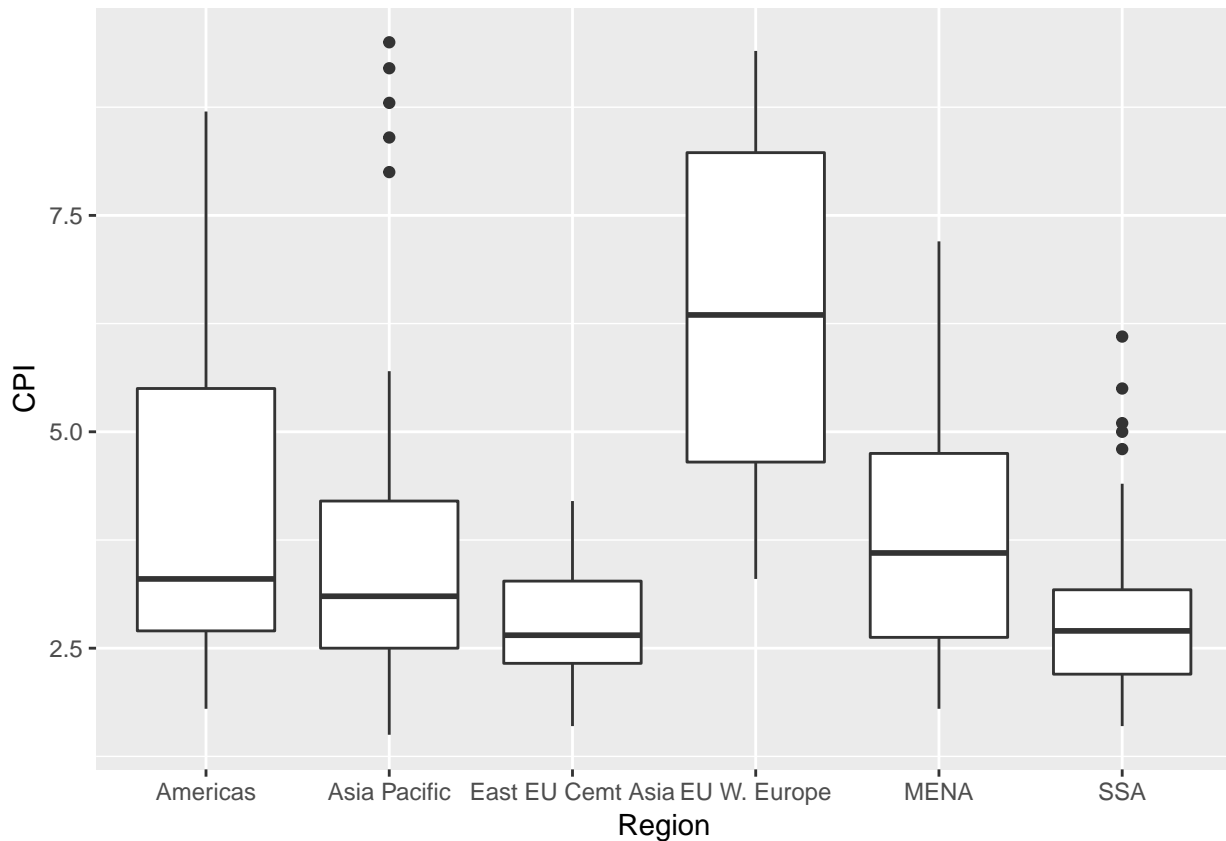
```
## Warning in `[<-.factor`(`*tmp*`, dat$Country == "Trinidad and Tobago",  
## value = structure(c(1L, : invalid factor level, NA generated
```

```
dat$Country[dat$Country=="Congo Republic"] = "Democratic Republic of the Congo"
```

```
## Warning in `[<-.factor`(`*tmp*`, dat$Country == "Congo Republic", value =  
## structure(c(1L, : invalid factor level, NA generated
```

## Corruption Perceptions

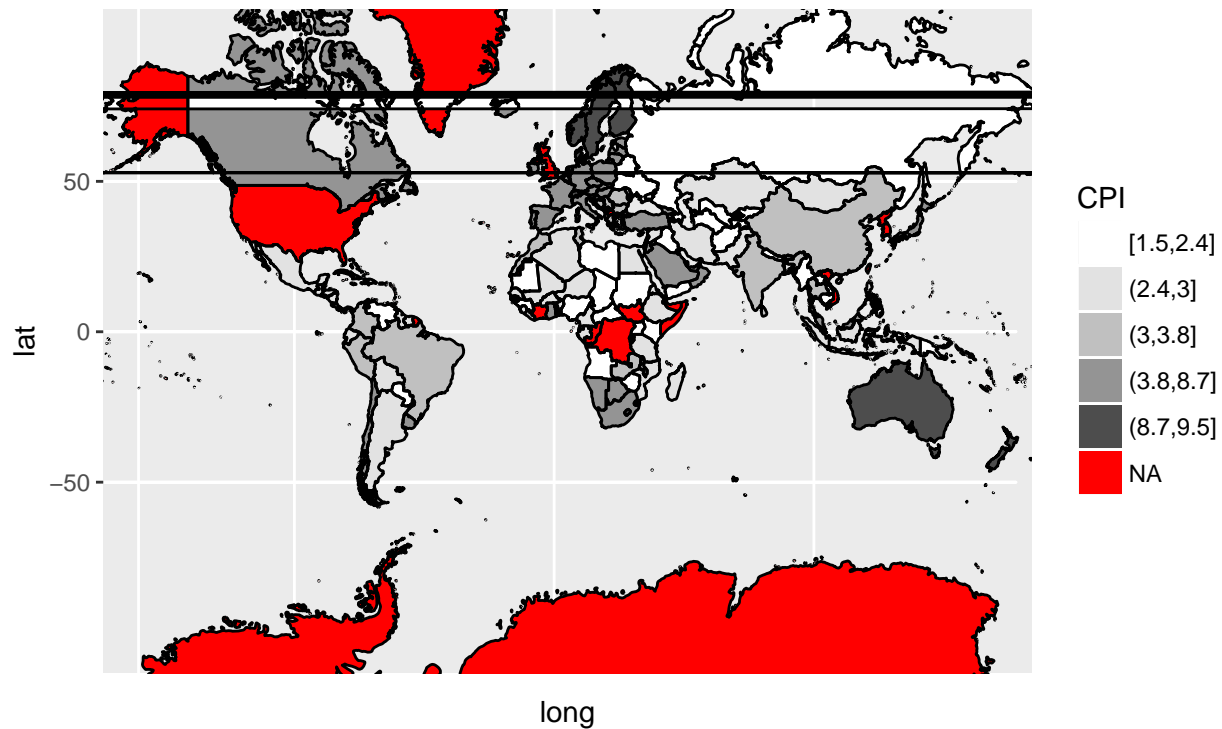
```
dat %>%  
  ggplot(aes(x = Region, y = CPI)) + geom_boxplot()
```



## Corruption Perceptions on the world map

```
map_data = merge(dfworldmap, dat, by.x="region", by.y="Country", all.x=TRUE)  
map_data = transform(map_data, CPI = cut_number(CPI, 5))  
map_data = map_data[order(map_data$order),]  
map_data %>%  
  ggplot(aes(x=long, y=lat, group=group)) +  
  geom_polygon(aes(fill = CPI)) +  
  scale_fill_grey(start=1, end=.3) +  
  geom_path(colour='black') +  
  coord_map() +  
  ggtitle("Corruption Perceptions Index Ranking of the world")
```

## Corruption Perceptions Index Ranking of the world

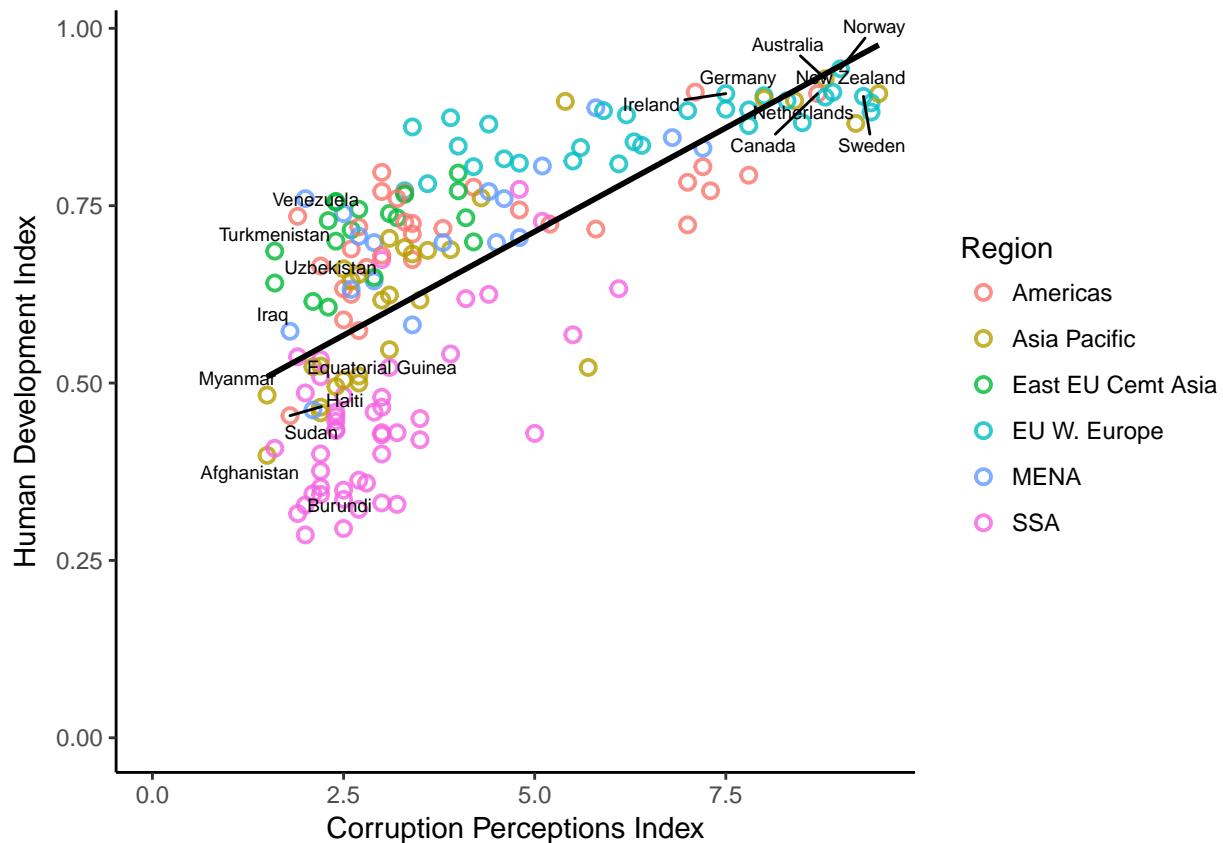


## Corruption Perceptions and Human Development

`geom_smooth()` is used to add a smooth line.

```
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() +
  geom_text_repel(data=subset(dat, HDI.Rank <=10 | CPI.Rank <=10), aes(CPI, HDI, label = Country), size=2)
  expand_limits(x = 0, y = 0)
```

## Warning: Removed 1 rows containing missing values (geom\_text\_repel).



## 2. Visualization on regional level

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)
reg_dat$CPI.Rank = rank(reg_dat$avgCPI, ties.method="first")
reg_dat$HDI.Rank = rank(reg_dat$avgHDI, ties.method="first")
```

### Corruption Perceptions and Human Development

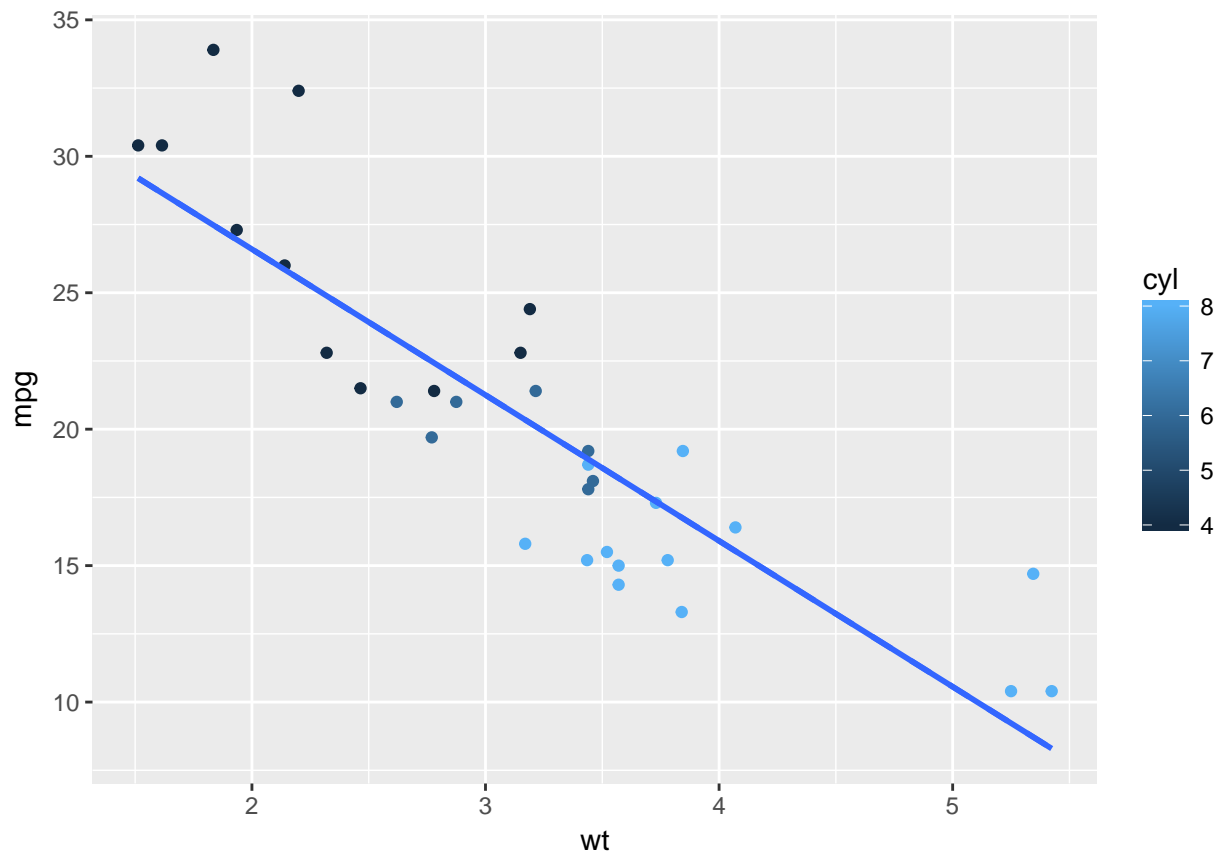
```
reg_dat %>%
  ggplot(aes(x = avgCPI, y = avgHDI)) +
  geom_point(aes(col=Region), size=5) +
  labs(x="Corruption Perceptions Index", y="Human Development Index") +
  theme_classic() +
  geom_text_repel(aes(avgCPI, avgHDI, label = Region), size=3) +
  expand_limits(x = 0, y = 0) +
  guides(col=F)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



Example: a lm for entire dataset and other lms for subsets

### Plot 3: include a lm for the entire dataset in its whole

```
ggplot(mtcars, aes(x = wt, y = mpg, col = cyl)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  geom_smooth(aes(group = 1), method = "lm", se = FALSE, linetype = 2)
```



Source:  
<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>