MindTheGap

Callum Hollywood

28 December 2017

## A Brief Exploratory Data Analysis of the Gapminder Dataset

## Shipped with the R Gapminder Package

#### We can get a feel for the general composition of this dataset

#### by looking at both ends of the dataframe, as it is shipped. Supported by some

#### additional basic descriptive statistics: …

head(gapminder)

## # A tibble: 6 x 6  
## country continent year lifeExp pop gdpPercap  
## <fctr> <fctr> <int> <dbl> <int> <dbl>  
## 1 Afghanistan Asia 1952 28.801 8425333 779.4453  
## 2 Afghanistan Asia 1957 30.332 9240934 820.8530  
## 3 Afghanistan Asia 1962 31.997 10267083 853.1007  
## 4 Afghanistan Asia 1967 34.020 11537966 836.1971  
## 5 Afghanistan Asia 1972 36.088 13079460 739.9811  
## 6 Afghanistan Asia 1977 38.438 14880372 786.1134

tail(gapminder)

## # A tibble: 6 x 6  
## country continent year lifeExp pop gdpPercap  
## <fctr> <fctr> <int> <dbl> <int> <dbl>  
## 1 Zimbabwe Africa 1982 60.363 7636524 788.8550  
## 2 Zimbabwe Africa 1987 62.351 9216418 706.1573  
## 3 Zimbabwe Africa 1992 60.377 10704340 693.4208  
## 4 Zimbabwe Africa 1997 46.809 11404948 792.4500  
## 5 Zimbabwe Africa 2002 39.989 11926563 672.0386  
## 6 Zimbabwe Africa 2007 43.487 12311143 469.7093

str(gapminder)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 1704 obs. of 6 variables:  
## $ country : Factor w/ 142 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ continent: Factor w/ 5 levels "Africa","Americas",..: 3 3 3 3 3 3 3 3 3 3 ...  
## $ year : int 1952 1957 1962 1967 1972 1977 1982 1987 1992 1997 ...  
## $ lifeExp : num 28.8 30.3 32 34 36.1 ...  
## $ pop : int 8425333 9240934 10267083 11537966 13079460 14880372 12881816 13867957 16317921 22227415 ...  
## $ gdpPercap: num 779 821 853 836 740 ...

summary(gapminder)

## country continent year lifeExp   
## Afghanistan: 12 Africa :624 Min. :1952 Min. :23.60   
## Albania : 12 Americas:300 1st Qu.:1966 1st Qu.:48.20   
## Algeria : 12 Asia :396 Median :1980 Median :60.71   
## Angola : 12 Europe :360 Mean :1980 Mean :59.47   
## Argentina : 12 Oceania : 24 3rd Qu.:1993 3rd Qu.:70.85   
## Australia : 12 Max. :2007 Max. :82.60   
## (Other) :1632   
## pop gdpPercap   
## Min. : 60011 Min. : 241.2   
## 1st Qu.: 2793664 1st Qu.: 1202.1   
## Median : 7023596 Median : 3531.8   
## Mean : 29601212 Mean : 7215.3   
## 3rd Qu.: 19585222 3rd Qu.: 9325.5   
## Max. :1318683096 Max. :113523.1   
##

#### Data Summary: –

#### It looks like we have over 1700 observations detailing measurements

#### for 3 key features, 1) Life Expectancy, 2) Population and 3) GDP per Capita

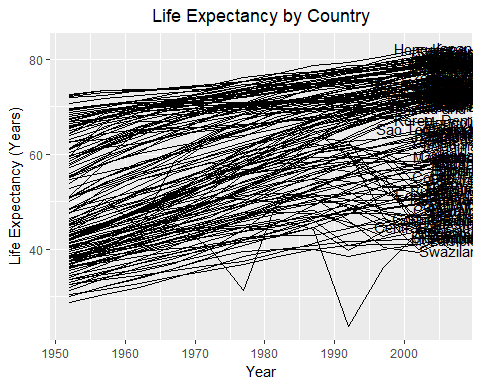
#### for 142 countries, which have also been labeled with their appropriate continent.

#### The duration for capturing this data appears to be from 1952 to 2007 and has been captured

#### at 5 year intervals.

#### We can proceed to examine the gapminder features, by country, over time: …

#========================  
# plot lifeExp over the survey duration, by country   
#========================  
  
ggplot(gapminder, aes(x = year, y = lifeExp, group = country)) +  
 geom\_line() +  
 theme(legend.position = "none") +  
 ggtitle("Life Expectancy by Country") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 geom\_text(data = gapminder[gapminder$year == 2007,], aes(label = country)) +  
 xlab("Year") +  
 ylab("Life Expectancy (Years)")



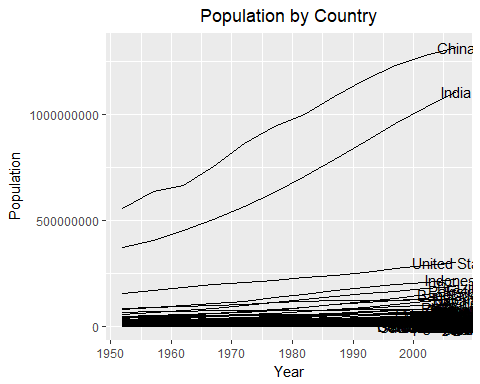
#### Looks like there’s a lot going on with all the countries represented above.

#### What we can discern though is that there appears to be a general increase in

#### life expectancy, though there are definitely some countries that seem to buck that trend.

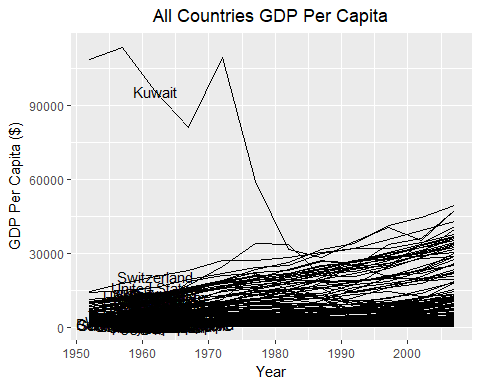
#### We will return to this and examine in greater depth!

#========================  
# plot pop over the survey duration, by country   
#========================  
  
ggplot(gapminder, aes(year, pop, group = country)) +  
 geom\_line() +  
 theme(legend.position = "none") +  
 ggtitle("Population by Country") +  
 geom\_text(data = gapminder[gapminder$year == 2007,], aes(label = country)) +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 xlab("Year") +  
 ylab("Population")



#### Jeez Louise - Look at the population climbs by China and India!

#========================  
# plot gdpPercap over the survey duration, by country   
#========================  
  
ggplot(gapminder, aes(year, gdpPercap, group = country)) +  
 geom\_line() +  
 theme(legend.position = "none") +  
 ggtitle("All Countries GDP Per Capita") +  
 geom\_text(data = gapminder[gapminder$year == 1962,], aes(label = country)) +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 xlab("Year") +  
 ylab("GDP Per Capita ($)")



#### In generel, i.e. omitting the GDP per Capita of Kuwait, we can see that this feature

#### tends to be increasing over time. It would be interesting to see what is going on with Kuwait later.

#### And if there are any other countries that have a reduced GDP

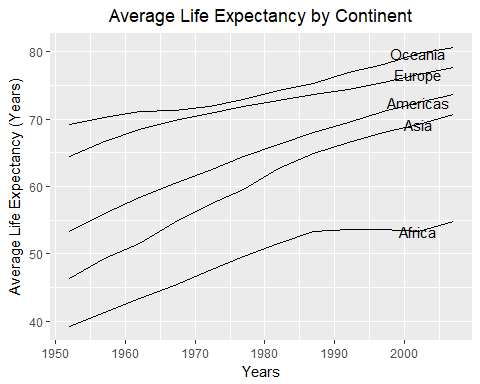
#### The graphs above show the gapminder features at the country level, and it all got

#### a bit busy on the eyes. Let’s aggregate to the ‘Continent’ level: …

continent\_view <- gapminder %>%   
 group\_by(continent, year) %>%   
 summarise(mean\_lifeExp = mean(lifeExp, na.rm = T),  
 mean\_pop = mean(pop, na.rm = T),  
 mean\_gdpPercap = mean(gdpPercap, na.rm = T))  
  
head(continent\_view)

## Source: local data frame [6 x 5]  
## Groups: continent [1]  
##   
## # A tibble: 6 x 5  
## continent year mean\_lifeExp mean\_pop mean\_gdpPercap  
## <fctr> <int> <dbl> <dbl> <dbl>  
## 1 Africa 1952 39.13550 4570010 1252.572  
## 2 Africa 1957 41.26635 5093033 1385.236  
## 3 Africa 1962 43.31944 5702247 1598.079  
## 4 Africa 1967 45.33454 6447875 2050.364  
## 5 Africa 1972 47.45094 7305376 2339.616  
## 6 Africa 1977 49.58042 8328097 2585.939

#========================  
# plot mean\_lifeExp over the survey duration, by continent   
#========================  
  
continent\_view %>%   
 select(-c(mean\_pop, mean\_gdpPercap)) %>%   
 ggplot(aes(year, mean\_lifeExp, group = continent)) +  
 geom\_line() +  
 ggtitle("Average Life Expectancy by Continent") +  
 geom\_text(data = continent\_view %>% filter(year == 2002) ,  
 aes(label = continent))+  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 xlab("Years") +  
 ylab("Average Life Expectancy (Years)")

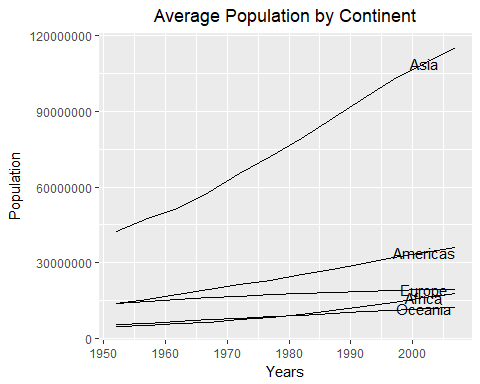


#### Across the board we can see that life expecnacy is on the increase and that there appears to

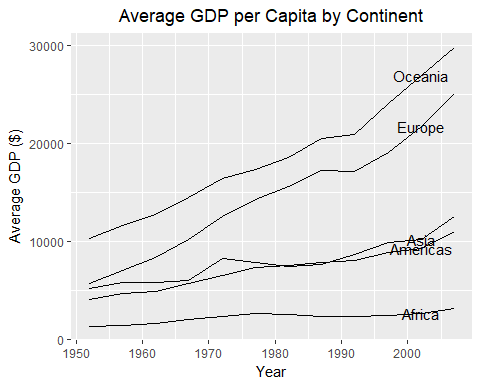
#### be a distinct difference in average life expectancy across the continents. Of particular note

#### is that not ony is Africa trailing, it appears to have levelled off through the late 80’s and 90’s.

#========================  
# plot mean\_pop over the survey duration, by continent   
#========================  
  
continent\_view %>%   
 select(-c(mean\_lifeExp, mean\_gdpPercap)) %>%   
 ggplot(aes(year, mean\_pop, group = continent)) +  
 geom\_line() +  
 ggtitle("Average Population by Continent") +  
 geom\_text(data = continent\_view %>% filter(year == 2002),  
 aes(label = continent)) +  
 xlab("Years") +  
 ylab("Population") +  
 theme(plot.title = element\_text(hjust = 0.5))



#========================  
# plot mean\_gdpPercap over the survey duration, by continent   
#========================  
  
continent\_view %>%   
 select(-c(mean\_lifeExp, mean\_pop)) %>%   
 ggplot(aes(year, mean\_gdpPercap, group = continent)) +  
 geom\_line() +  
 ggtitle("Average GDP per Capita by Continent") +  
 geom\_text(data = continent\_view %>% filter(year == 2002),  
 aes(label = continent)) +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 xlab("Year") +  
 ylab("Average GDP ($)")



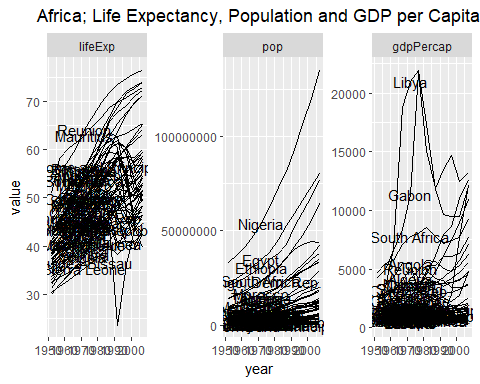
#### It appears that the average GDP in both Oceania and Eurpoe is outpacing

#### Asia and the Americas, which are all pulling away from Africa

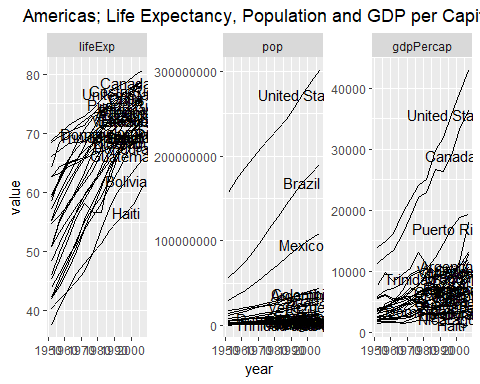
#### Africa definitely seems to out of step with the other continents.

#### Let’s have a look at the continents individually to see how the gapminder features are interplaying:…

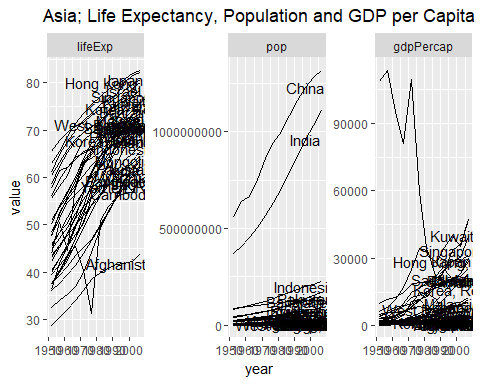
africa <- gapminder %>% filter(continent == "Africa") %>% select(-c(continent))  
  
africa\_melt <- melt(africa,   
 id.vars = c("country", "year"),  
 variable.name = "feature",  
 value.name = "value")   
  
africa\_melt %>%   
 ggplot(aes(year, value, group = country)) +  
 geom\_line() +  
 facet\_wrap(~ feature, scales = "free\_y") +  
 geom\_text(data = africa\_melt %>% filter(year == 1972),  
 aes(label = country)) +  
 ggtitle("Africa; Life Expectancy, Population and GDP per Capita") +  
 theme(plot.title = element\_text(hjust = 0.5))



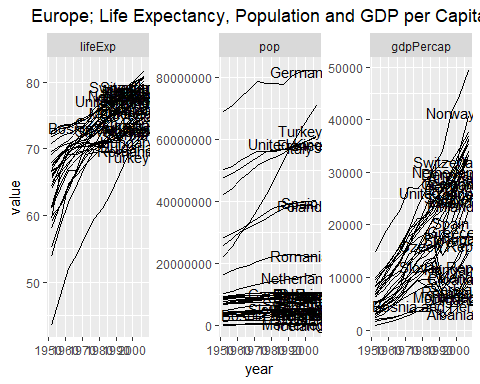
americas <- gapminder %>% filter(continent == "Americas") %>% select(-c(continent))  
  
americas\_melt <- melt(americas,   
 id.vars = c("country", "year"),  
 variable.name = 'feature',  
 value.name = "value")  
  
americas\_melt %>%   
 ggplot(aes(year, value, group = country)) +  
 geom\_line() +  
 facet\_wrap(~feature, scales = 'free\_y') +  
 geom\_text(data = americas\_melt %>% filter(year == 1997),  
 aes(label = country)) +  
 ggtitle("Americas; Life Expectancy, Population and GDP per Capita") +  
 theme(plot.title = element\_text(hjust = 0.5))



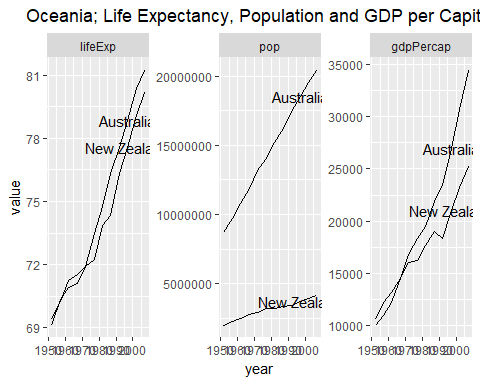
asia <- gapminder %>% filter(continent == "Asia") %>% select(-c(continent))  
  
asia\_melt <- melt(asia,   
 id.vars = c("country", "year"),  
 variable.name = "feature",  
 value.name = "value")  
  
asia\_melt %>%   
 ggplot(aes(year, value, group = country)) +  
 geom\_line() +  
 facet\_wrap(~ feature, scales = "free\_y") +  
 geom\_text(data = asia\_melt %>% filter(year == 1997),  
 aes(label = country)) +  
 ggtitle("Asia; Life Expectancy, Population and GDP per Capita") +  
 theme(plot.title = element\_text(hjust = 0.5))



europe <- gapminder %>% filter(continent == "Europe") %>% select(-c(continent))  
  
europe\_melt <- melt(europe,  
 id.vars = c("country", "year"),  
 variable.name = "feature",  
 value.name = "value")  
  
  
europe\_melt %>%   
 ggplot(aes(year, value, group = country)) +  
 geom\_line() +  
 facet\_wrap(~ feature, scales = 'free\_y') +  
 geom\_text(data = europe\_melt %>% filter(year == 1997),  
 aes(label = country)) +  
 ggtitle("Europe; Life Expectancy, Population and GDP per Capita") +  
 theme(plot.title = element\_text(hjust = 0.5))

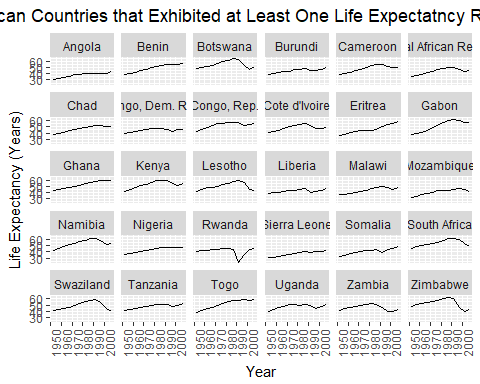


oceania <- gapminder %>% filter(continent == "Oceania") %>% select(-c(continent))  
  
oceania\_melt <- melt(oceania,  
 id.vars = c("country", "year"),  
 variable.name = "feature",  
 value.name = "value")  
  
oceania\_melt %>%   
 ggplot(aes(year, value, group = country)) +  
 geom\_line() +  
 facet\_wrap(~ feature, scales = "free\_y") +  
 geom\_text(data = oceania\_melt %>% filter(year == 1997),  
 aes(label = country)) +  
 ggtitle("Oceania; Life Expectancy, Population and GDP per Capita") +  
 theme(plot.title = element\_text(hjust = 0.5))



#### Africa still requires further examination. Let’s have a look at all the countries where there was a dip in life expectancy observed, at some/any point in the data:…

##################################################################  
# Drill into Africa and examine at the country level  
# Filter out countries that never took a dip in lifeExp at any point  
# Facet for better visuals  
##################################################################  
  
lifeexpdrops <- gapminder::gapminder %>%   
 filter(continent == "Africa") %>%   
 select(-c(continent, pop, gdpPercap)) %>%   
 spread(year, lifeExp)  
  
indx <- colSums(apply(lifeexpdrops %>% select(-c(country)), 1L, diff) > 0L) == (ncol(lifeexpdrops%>% select(-c(country))) - 1L)  
lifeexpdrops <- lifeexpdrops[!indx,]$country  
  
gapminder::gapminder %>%   
 filter(continent == "Africa") %>%   
 filter(country %in% lifeexpdrops) %>%   
 select(-c(continent)) %>%   
 ggplot(aes(year, lifeExp, group = country)) +  
 geom\_line() +  
 facet\_wrap(~ country) +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1)) +  
 ggtitle("African Countries that Exhibited at Least One Life Expectatncy Reduction") +  
 theme(plot.title = element\_text(hjust = 0.5)) +  
 xlab("Year") +  
 ylab("Life Expectancy (Years)")

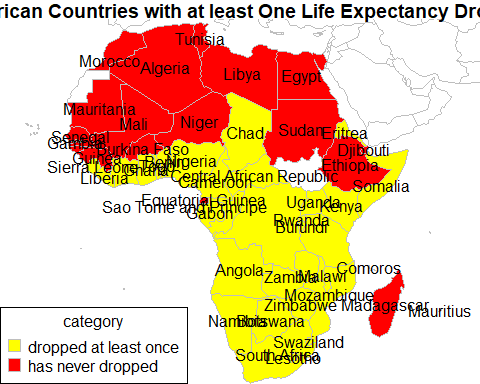


#### Let’s start to visualise with a map or two:…

lifeexpdrops <- gapminder::gapminder %>%   
 filter(continent == "Africa") %>%   
 select(-c(continent, pop, gdpPercap)) %>%   
 spread(year, lifeExp)  
  
indx <- colSums(apply(lifeexpdrops %>% select(-c(country)), 1L, diff) > 0L) == (ncol(lifeexpdrops%>% select(-c(country))) - 1L)  
lifeexpdrops <- lifeexpdrops[!indx,]$country  
  
africa <- gapminder::gapminder %>%   
 filter(continent == "Africa") %>%   
 select(-c(continent)) %>%   
 filter(year == 2002) %>%   
 mutate(lifeExpDrop = if\_else(country %in% lifeexpdrops, "dropped at least once", "has never dropped"))  
  
  
sPDF <- joinCountryData2Map(africa,  
 joinCode = "NAME",  
 nameJoinColumn = "country")

## 51 codes from your data successfully matched countries in the map  
## 1 codes from your data failed to match with a country code in the map  
## 192 codes from the map weren't represented in your data

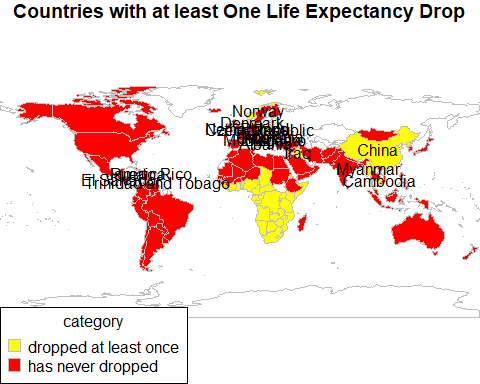
par(mai = c(0,0,0.2,0), xaxs="i", yaxs = "i")  
mapCountryData(sPDF,   
 nameColumnToPlot = "lifeExpDrop",   
 catMethod="categorical",   
 numCats = 2,  
 addLegend =T,  
 mapTitle = "African Countries with at least One Life Expectancy Drop",  
 mapRegion = "Africa")  
  
country\_coord<-data.frame(coordinates(sPDF),stringsAsFactors=F)  
country\_coord\_B <- country\_coord[row.names(country\_coord) %in% africa$country,]  
  
text(x = country\_coord\_B$X1, y = country\_coord\_B$X2,   
 labels = row.names(country\_coord\_B))



lifeexpdrops <- gapminder::gapminder %>%   
 select(-c(continent, pop, gdpPercap)) %>%   
 spread(year, lifeExp)  
  
indx <- colSums(apply(lifeexpdrops %>% select(-c(country)), 1L, diff) > 0L) == (ncol(lifeexpdrops%>% select(-c(country))) - 1L)  
lifeexpdrops <- lifeexpdrops[!indx,]$country  
  
worldLifeExpDrop <- gapminder::gapminder %>%   
 select(-c(continent)) %>%   
 filter(year == 2002) %>%   
 mutate(lifeExpDrop = if\_else(country %in% lifeexpdrops, "dropped at least once", "has never dropped"))  
  
  
worldLifeExpDropExcptAfrica <- worldLifeExpDrop %>%   
 filter(lifeExpDrop == "dropped at least once") %>%   
 filter(!country %in% africa$country)  
  
sPDF\_World <- joinCountryData2Map(worldLifeExpDrop,  
 joinCode = "NAME",  
 nameJoinColumn = "country")

## 139 codes from your data successfully matched countries in the map  
## 3 codes from your data failed to match with a country code in the map  
## 104 codes from the map weren't represented in your data

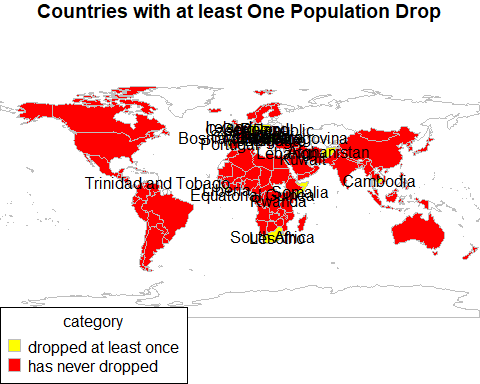
par(mai = c(0,0,0.2,0), xaxs="i", yaxs = "i")  
mapCountryData(sPDF\_World,   
 nameColumnToPlot = "lifeExpDrop",   
 catMethod="categorical",   
 numCats = 2,  
 addLegend =T,  
 mapTitle = "Countries with at least One Life Expectancy Drop"  
 )  
  
country\_coord <- data.frame(coordinates(sPDF\_World),stringsAsFactors=F)  
country\_coord\_C <- country\_coord[row.names(country\_coord) %in% worldLifeExpDropExcptAfrica$country,]  
  
text(x = country\_coord\_C$X1, y = country\_coord\_C$X2,   
 labels = row.names(country\_coord\_C))



pop\_drops <- gapminder::gapminder %>%   
 select(-c(continent, lifeExp, gdpPercap)) %>%   
 spread(year, pop)  
  
indx <- colSums(apply(pop\_drops %>% select(-c(country)), 1L, diff) > 0L) == (ncol(pop\_drops%>% select(-c(country))) - 1L)  
pop\_drops <- pop\_drops[!indx,]$country  
  
world\_popdrops <- gapminder::gapminder %>%   
 select(-c(continent)) %>%   
 filter(year == 2002) %>%   
 mutate(pop\_drops = if\_else(country %in% pop\_drops, "dropped at least once", "has never dropped"))  
  
world\_popdrops\_slim <- world\_popdrops[world\_popdrops$pop\_drops == 'dropped at least once',]  
  
  
  
  
sPDF\_worldpopdrop <- joinCountryData2Map(world\_popdrops,  
 joinCode = "NAME",  
 nameJoinColumn = "country")

## 139 codes from your data successfully matched countries in the map  
## 3 codes from your data failed to match with a country code in the map  
## 104 codes from the map weren't represented in your data

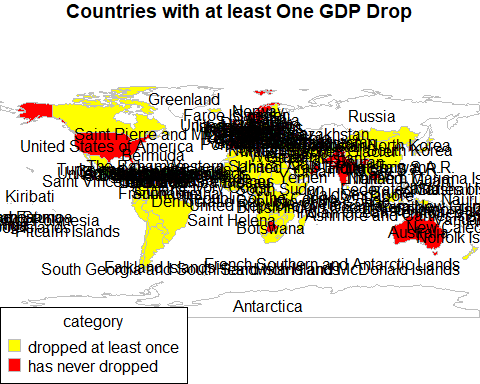
par(mai = c(0,0,0.2,0), xaxs="i", yaxs = "i")  
mapCountryData(sPDF\_worldpopdrop,   
 nameColumnToPlot = "pop\_drops",   
 catMethod="categorical",   
 numCats = 2,  
 addLegend =T,  
 mapTitle = "Countries with at least One Population Drop")  
  
country\_coord <- data.frame(coordinates(sPDF\_World),stringsAsFactors=F)  
country\_coord\_D <- country\_coord[row.names(country\_coord) %in% world\_popdrops\_slim$country,]  
  
text(x = country\_coord\_D$X1, y = country\_coord\_D$X2,   
 labels = row.names(country\_coord\_D))



gdp\_drops <- gapminder::gapminder %>%   
 select(-c(continent, lifeExp, pop)) %>%   
 spread(year, gdpPercap)  
  
indx <- colSums(apply(gdp\_drops %>% select(-c(country)), 1L, diff) > 0L) == (ncol(gdp\_drops%>% select(-c(country))) - 1L)  
gdp\_drops <- gdp\_drops[!indx,]$country  
  
  
  
  
world\_gdpdrops <- gapminder::gapminder %>%   
 select(-c(continent)) %>%   
 filter(year == 2002) %>%   
 mutate(gdp\_drops = if\_else(country %in% gdp\_drops, "dropped at least once", "has never dropped"))  
  
world\_gdpdrops\_slim <- world\_gdpdrops[world\_gdpdrops$gdp\_drops == 'dropped at least once',]  
  
  
sPDF\_worldgdpdrop <- joinCountryData2Map(world\_gdpdrops,  
 joinCode = "NAME",  
 nameJoinColumn = "country")

## 139 codes from your data successfully matched countries in the map  
## 3 codes from your data failed to match with a country code in the map  
## 104 codes from the map weren't represented in your data

par(mai = c(0,0,0.2,0), xaxs="i", yaxs = "i")  
mapCountryData(sPDF\_worldgdpdrop,   
 nameColumnToPlot = "gdp\_drops",   
 catMethod="categorical",   
 numCats = 2,  
 addLegend =T,  
 mapTitle = "Countries with at least One GDP Drop")  
  
country\_coord <- data.frame(coordinates(sPDF\_World),stringsAsFactors=F)  
country\_coord\_E <- country\_coord[!row.names(country\_coord) %in% world\_gdpdrops\_slim$country,]  
  
text(x = country\_coord\_E$X1, y = country\_coord\_E$X2,   
 labels = row.names(country\_coord\_E))



#### It appears that with respect to GDP it is not unusual to have a dip in GDP year on year,

#### for many countries. Let’s instead have a look to see if any countries have GDPs in 2007 that

#### are less than our starting point in 1952: …

gdp\_world\_drops <- gapminder::gapminder %>%   
 filter(year %in% c(1952, 2007)) %>%   
 select(-c(lifeExp, pop, continent)) %>%   
 spread(year, gdpPercap) %>%   
 mutate(gdp\_reduced = `2007` - `1952`) %>%   
 mutate(gdp\_reduced\_cat = if\_else(gdp\_reduced < 0, "Has Reduced Overall", "Has Increased Overall")) %>%   
 filter(gdp\_reduced\_cat == "Has Reduced Overall")  
  
gdp\_world\_drops %>% select(-c(gdp\_reduced\_cat))

## # A tibble: 12 x 4  
## country `1952` `2007` gdp\_reduced  
## <fctr> <dbl> <dbl> <dbl>  
## 1 Central African Republic 1071.3107 706.0165 -365.29418  
## 2 Comoros 1102.9909 986.1479 -116.84306  
## 3 Congo, Dem. Rep. 780.5423 277.5519 -502.99047  
## 4 Djibouti 2669.5295 2082.4816 -587.04791  
## 5 Haiti 1840.3669 1201.6372 -638.72978  
## 6 Kuwait 108382.3529 47306.9898 -61075.36312  
## 7 Liberia 575.5730 414.5073 -161.06565  
## 8 Madagascar 1443.0117 1044.7701 -398.24159  
## 9 Nicaragua 3112.3639 2749.3210 -363.04298  
## 10 Niger 761.8794 619.6769 -142.20248  
## 11 Sierra Leone 879.7877 862.5408 -17.24698  
## 12 Somalia 1135.7498 926.1411 -209.60877

sPDF\_gdp\_world\_drops <- joinCountryData2Map(gdp\_world\_drops,  
 joinCode = "NAME",  
 nameJoinColumn = "country")

## 12 codes from your data successfully matched countries in the map  
## 0 codes from your data failed to match with a country code in the map  
## 231 codes from the map weren't represented in your data

par(mai = c(0,0,0.2,0), xaxs="i", yaxs = "i")  
mapCountryData(sPDF\_gdp\_world\_drops,   
 nameColumnToPlot = "gdp\_reduced\_cat",   
 catMethod="categorical",   
 numCats = 2,  
 addLegend =F,  
 mapTitle = "Countries with a Smaller GDP in 2007 than in 1952"  
)  
  
country\_coord <- data.frame(coordinates(sPDF\_gdp\_world\_drops),stringsAsFactors=F)  
country\_coord\_F <- country\_coord[row.names(country\_coord) %in% sPDF\_gdp\_world\_drops$country,]  
  
text(x = country\_coord\_F$X1, y = country\_coord\_F$X2,   
 labels = row.names(country\_coord\_F))

