Data Wrangling

Patrick Kelly

2/10/2019

Data shaping and cleaning was required before the 2 datasets could be merged.

```
library(data.table)
library(dplyr)
```

```
tb_data <- fread("https://raw.githubusercontent.com/datamustekeers/WHOtb_data_analysis/master/data/TB_bu
names(tb_data) <- c("Country", "year","population","tb_cases")
tb_data$population <- round(tb_data$population/1e+6,2)
tb_data$incidence <- round((tb_data$tb_cases*0.1)/tb_data$population,1)
all_countries = unique(tb_data$Country)
all_countries <- data.frame(Country=all_countries)
head(all_countries)
```

```
##          Country
## 1    Afghanistan
## 2        Albania
## 3        Algeria
## 4 American Samoa
## 5        Andorra
## 6         Angola
```

Load the Country and ISO Codes

```
world <- fread("https://raw.githubusercontent.com/damonzon/WHO_TB_Burden/master/world2.csv")
names(world)
```

```
## [1] "Country"             "alpha-2"             "alpha-3"
## [4] "sub-region"          "intermediate-region" "region-code"
```

```
names(world) <- c("Country", "alpha_2","alpha_3",
    "sub_region","intermediate_region","region_code")
names(world)
```

```
## [1] "Country"             "alpha_2"             "alpha_3"
## [4] "sub_region"          "intermediate_region" "region_code"
```

```
class(world$Country)
```

```
## [1] "character"
```

Harmonize country names tetween the 2 datasets

```
world$Country[28] <- "Bonaire, Saint Eustatius and Saba"
world$Country[52] <- "Democratic Republic of the Congo"
world$Country[71] <- "Swaziland" # Eswatini
world$Country[119] <- "Democratic People's Republic of Korea"
world$Country[120] <- "Republic of Korea"
world$Country[133] <- "The Former Yugoslav Republic of Macedonia"
world$Country[147] <- "Republic of Moldova"
world$Country[171] <- "West Bank and Gaza Strip"
world$Country[220] <- "United Republic of Tanzania"
world$Country[243] <- "British Virgin Islands"
world$Country[245] <- "Wallis and Futuna Islands"
world <- arrange(world,Country)
```

Merge the datasets

```
class(all_countries$Country)
```

```
## [1] "factor"
```

```
all_countries$Country <- as.character(all_countries$Country)
class(all_countries$Country)
```

```
## [1] "character"
```

```
data <- left_join(all_countries,world, by = "Country")
regions <- merge(tb_data,data, by = "Country")
names(regions)
```

```
##  [1] "Country"             "year"                "population"
##  [4] "tb_cases"            "incidence"           "alpha_2"
##  [7] "alpha_3"             "sub_region"          "intermediate_region"
## [10] "region_code"
```

```
regions$region_code <- ifelse(regions$region_code=="",
      regions$intermediate_region,regions$region_code)
regions$intermediate_region <-NULL
names(regions)
```

```
## [1] "Country"     "year"        "population"  "tb_cases"    "incidence"
## [6] "alpha_2"     "alpha_3"     "sub_region"  "region_code"
```

```
# write.csv(regions,"tb_regions.csv",row.names=FALSE)
```