Malaria Dataset Analysis

This document contains my malaria dataset analysis.

Dataset Used:
https://github.com/rfordatascience/tidytuesday/tree/master/data/2018/2018-11-13

Import relevant libraries

```
#import libraries
#if you havent install the relevant package, please install them.
#install.packages('dplyr')
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

#install.packages('ggplot2')
library(ggplot2)
```

Malaria_inc dataset analysis

Answer I would like to find out from this dataset analysis:

1. Which countries have the highest malaria cases? (top 10)
2. From the top 10 countries from (1), are there any improvement over the years?

```
#reading malaria_incidence dataset
malaria_inc <-read.csv("malaria_inc.csv")
names(malaria_inc)

## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4]
"Incidence.of.malaria..per.1.000.population.at.risk...per.1.000.population.at
.risk."

str(malaria_inc)
```

```
## 'data.frame':    508 obs. of  4 variables:
##  $ Entity
: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Code
: chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ Year
: int  2000 2005 2010 2015 2000 2005 2010 2015 2000 2005 ...
##  $
Incidence.of.malaria..per.1.000.population.at.risk...per.1.000.population.at.
risk.: num  107.1 46.5 23.9 23.6 0.0377 ...
```

Observation:

1) The dataset only contains 4 years - 2000,2005,2010,2015. To keep the analysis consistent, I will use only these 4 values for the rest of my analysis.

2) The first column name is called "Entity" which is basically the countries. The last column name is "Incidence.of.malaria..per.1.000.population.at.risk...per.1.000.population.at.risk." which is the number of cases per 1000 population. (For better readibility, I will rename the two columns.)

```
#rename first column to country and last column to cases for better
readibility
malaria_inc<-malaria_inc %>% setNames(c("country", "code", "year", "cases"))

#find out the top 10 countries that have the highest cases for the span of 15
years
worst_inc <- malaria_inc %>%
  group_by(country) %>%
  summarise(cases = sum(cases)) %>%
  arrange(desc(cases))%>%
  head(10)
```

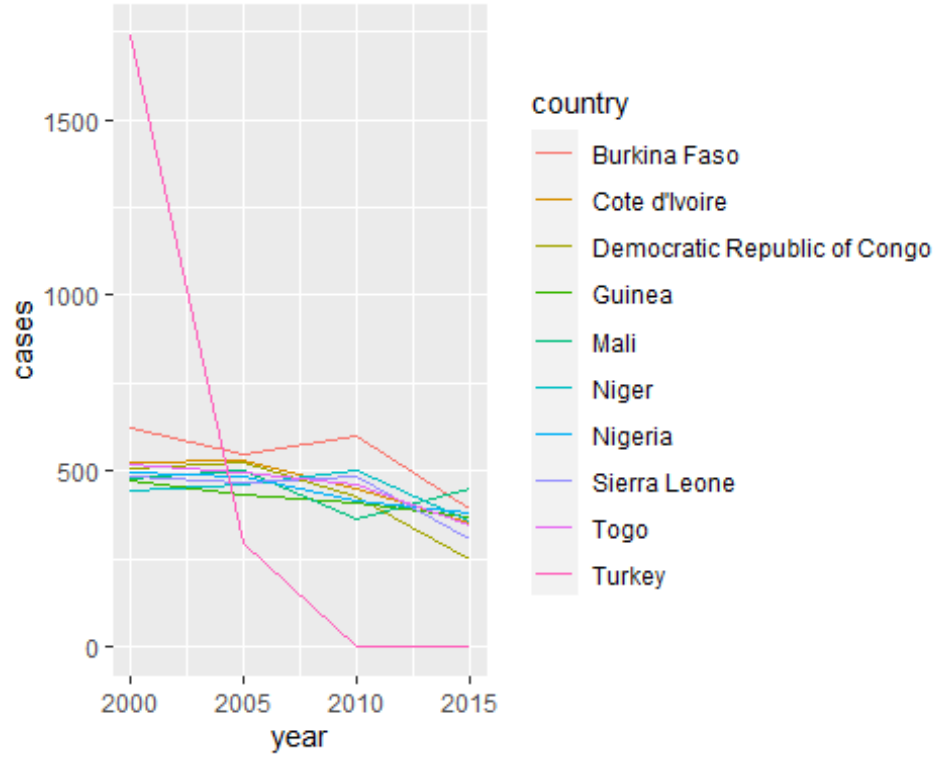
Observation:

The top 10 countries with the highest cases for the span of 15 years are Burkina Faso,Turkey, Cote d'Ivoire, Togo, Mali, Nigeria, Niger, Sierra Leone, Democratic Republic of Congo and Guinea.

```
#filtered dataset based on the the top 10 countries with the highest cases
worst_inc_year<-malaria_inc %>%
  group_by(year,country) %>%
  summarise(cases = sum(cases)) %>%
  arrange(country,year) %>%
  filter(country %in% worst_inc$country)

## `summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.
```

```
#plot to check whether there are any improvements over the years-decrease in cases
ggplot(data=worst_inc_year,aes(x=year,y=cases,color = country))+geom_line()
```



Observation:

For most of the countries, there were not much significant improvements in the reduction of cases. The only country that have significant improvement is Turkey.

Malaria_death dataset analysis

Answer I would like to find out from this dataset analysis:

1. Which countries have the highest malaria deaths? (top 10) Are the results the same as those countries with the highest cases?

2. From the top 10 countries from (1), are there any improvement over the years?

```
#reading malaria_deaths dataset
malaria_deaths <-read.csv("malaria_deaths.csv")
names(malaria_deaths)

## [1] "Entity"
## [2] "Code"
## [3] "Year"
## [4]
```

```
"Deaths...Malaria...Sex..Both...Age..Age.standardized..Rate...per.100.000.peo
ple."

str(malaria_deaths)

## 'data.frame':    6156 obs. of  4 variables:
##  $ Entity
: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Code
: chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ Year
: int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
##  $
Deaths...Malaria...Sex..Both...Age..Age.standardized..Rate...per.100.000.peop
le.: num  6.8 6.97 6.99 7.09 7.39 ...
```

Observations:

1) The first column name is called “Entity” which is basically the countries. The last column name is “Deaths...Malaria...Sex..Both...Age..Age.standardized..Rate...per.100.000.people.” which is the number of deaths per 100,000 people.(For better readibility, I will rename the two columns.)

```
#rename first column to country and last column to deaths for better
readibility
malaria_deaths<-malaria_deaths %>% setNames(c("country", "code", "year",
"deaths"))

#filter dataset to only contains Years = 2000,2005,2010,2015 to ensure the
analysis is consistent for comparison
malaria_deaths_filtered_year <- malaria_deaths %>%
  filter(year==2000|year == 2005|year==2010|year==2015)

#find out which countries have the worst death rates
worst_countries <- malaria_deaths_filtered_year %>%
  group_by(country) %>%
  summarise(deaths = sum(deaths)) %>%
  arrange(desc(deaths))%>%
  head(10)
```

Observations:

Comparing the top 10 countries with the highest death rates and highest case rates, 8 out of 10 are the same.This proves that with higher cases rate, there will be higher deaths.

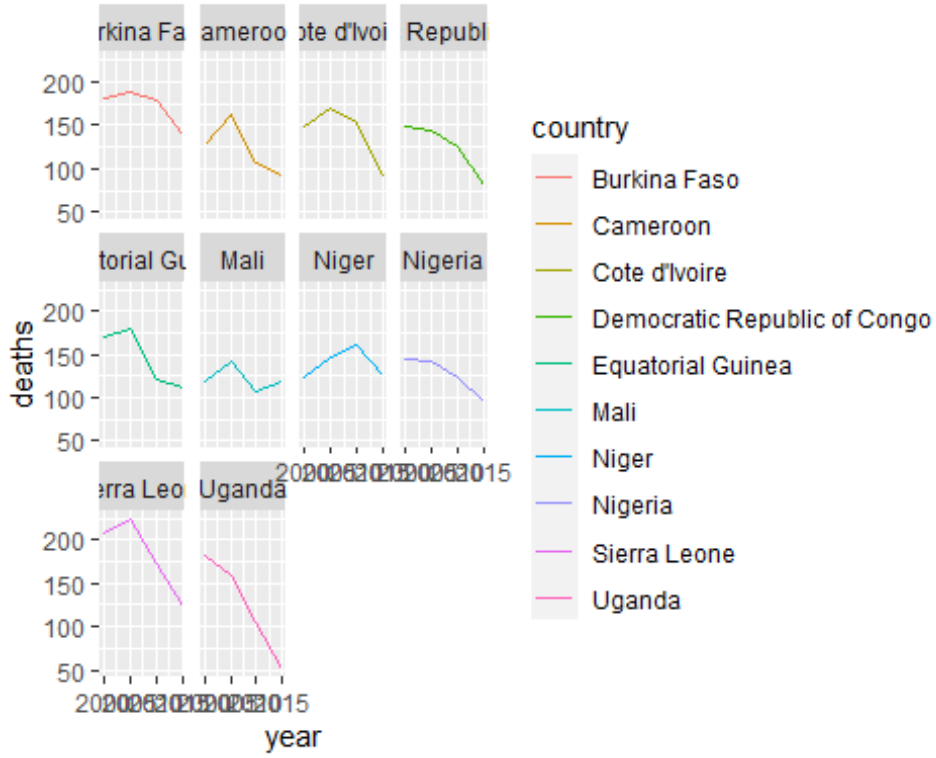
The top 10 countries with the highest death rates are Sierra Leone, Burkina Faso,Equatorial Guinea, Cote d’Ivoire, Niger, Nigeria, Uganda, Democratic Republic of Congo, Mali, Cameroon

```
#filtered dataset based on the the top 10 countries with the highest deaths
rate
```

```
worst_countries_year<-malaria_deaths_filtered_year %>%
  group_by(year,country) %>%
  summarise(deaths = sum(deaths)) %>%
  arrange(country,year) %>%
  filter(country %in% worst_countries$country)

## `summarise()` has grouped output by 'year'. You can override using the
`.groups` argument.

#plot to check whether there are decrement in death rates over the years
ggplot(data=worst_countries_year,aes(x=year,y=deaths,color =
country))+geom_line() +facet_wrap(~country)
```



Observation:

Most of the countries have a downward trends meaning over the span of 15 years, the death rates have been decreasing. There is only 1 country that have insignificant improvement - Mali.

Malaria_death_age dataset analysis

Answer I would like to find out from this dataset analysis:

1. For the top 10 countries with the highest death rates, which age group have a higher percentage of deaths?

```
#reading malaria_deaths_age dataset
malaria_deaths_age <-read.csv("malaria_deaths_age.csv")
names(malaria_deaths_age)
```

```
## [1] "X"         "entity"    "code"      "year"      "age_group" "deaths"
```

```
str(malaria_deaths_age)
```

```
## 'data.frame':    30780 obs. of  6 variables:
##  $ X        : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ entity   : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan"
...
##  $ code     : chr  "AFG" "AFG" "AFG" "AFG" ...
##  $ year     : int  1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 ...
##  $ age_group: chr  "Under 5" "Under 5" "Under 5" "Under 5" ...
##  $ deaths   : num  185 192 197 207 226 ...
```

Observations:

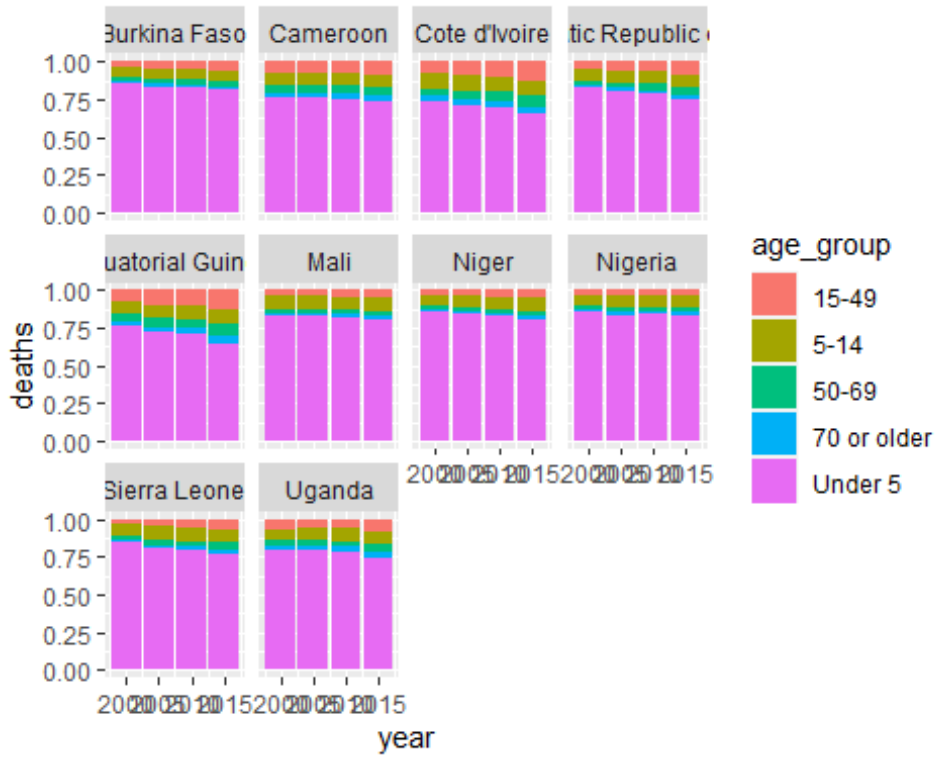
1) The first column name is called “x” which is the running index. The second column name is called “Entity” which is basically the countries.(For better readibility, I will rename the two columns.)

```
#rename the columns
malaria_deaths_age<-malaria_deaths_age %>% setNames(c("index","country",
"code", "year","age_group", "deaths"))

#filter dataset to only contains Years = 2000,2005,2010,2015 to ensure the
analysis is consistent for comparison
malaria_deaths_age_filtered_year<-malaria_deaths_age %>%
  filter(year==2000|year == 2005|year==2010|year==2015)

#filter based on the top 10 countries with the highest death rate.As this
dataset is related to death rates, I used the top 10 countries that has the
highest death rates(as produced by the second analysis).
worst_countries_year_age<-malaria_deaths_age_filtered_year %>%
  filter(country %in% worst_countries$country)

#plot to check the age group that are more vulnerable
ggplot(worst_countries_year_age, aes(fill=age_group, y=deaths, x=year)) +
    geom_bar(position="fill", stat="identity")+ facet_wrap(~country)
```



Observations:

For all the 10 countries with the highest death rates, the highest death rates age group is "Under 5". Therefore, it is important to focus any malaria treatment to children to reduce the death rates as children immunity is lower.

Summary

1) Poorly developed countries like Africa have the highest chance to suffer from Malaria.

2) The most vulnerable age group is children under 5.

3) Despite many efforts to help reduce the spread of Malaria, poorly developed countries are still suffering from Malaria although there is an significant improvement to the death rates.