Exercise: **Aggregation**

Day 2, Part C

```
> library(reshape2)
```

1. Load in the Ebola deaths data for West Africa (‘data/ebola_fatalities_sex_country.csv’).

```
> main_dir <- "C:/Users/ngraetz/Documents/repos/r_training_penn/"  # CHANGE TO YOUR LOCAL COPY O
> data <- read.csv(paste0(main_dir, "data/ebola_fatalities_sex_country.csv"), stringsAsFactors =
```

a. Using the data loaded in (1), create the data frame below reporting the number of deaths by country and gender:

```
> dcast(data, Country ~ Gender, value.var = "Deaths", fun.aggregate = sum)
Country Female   Male
1       Guinea 1001.9  930.1
2      Liberia 1002.4 1055.3
3 Sierra Leone 3140.0 2987.7
```

b. Using the data loaded in (1), create the data frame below reporting the number of deaths by age and country:

```
> dcast(data, Age ~ Country, value.var = "Deaths", fun.aggregate = sum)
Age Guinea Liberia Sierra Leone
1    0   46.4    14.4         85.9
2    1  115.5    95.9        354.1
3    5   89.8   102.4        368.5
4   10   63.3    97.1        383.9
5   15  108.9   128.5        410.1
6   20  146.4   174.2        554.2
7   25  208.8   222.6        691.8
8   30  195.1   258.2        578.0
9   35  201.4   237.0        629.5
10  40  160.1   224.2        496.1
11  45  150.9   183.0        415.1
12  50  116.1   128.9        310.6
13  55   92.0    63.5        207.5
14  60  100.9    50.1        218.3
15  65   43.8    26.0        170.5
16  70   52.8    25.0         78.1
17  75   15.9     8.0         64.4
18  80   23.9    18.7        111.1
```

c. Using the data loaded in (1), calculate the total number of fatalities by country, i.e.:

```
> country_deaths <- dcast(data, Country ~ ., value.var = "Deaths", fun.aggregate = sum)
> names(country_deaths)[2] <- "Deaths"
> country_deaths
Country Deaths
1       Guinea 1932.0
2      Liberia 2057.7
3 Sierra Leone 6127.7
```

d. Using the data loaded in (1), calculate the total number of fatalities by age, i.e.:

```
> age_deaths <- dcast(data, Age ~ ., value.var = "Deaths", fun.aggregate = sum)
> names(age_deaths)[2] <- "Deaths"
> age_deaths
Age Deaths
1    0  146.7
2    1  565.5
3    5  560.7
4   10  544.3
5   15  647.5
6   20  874.8
7   25 1123.2
8   30 1031.3
9   35 1067.9
10  40  880.4
11  45  749.0
12  50  555.6
13  55  363.0
14  60  369.3
15  65  240.3
16  70  155.9
17  75   88.3
18  80  153.7
```

e. Remove all of the data frames used in this question from your work space.

```
> rm(data, country_deaths, age_deaths)
```

Bonus:

2. Still using the original data set (‘data/ebola_fatalities_sex_country.csv’):

a. Find and read the help docs for `aggregate` and `apply`.

```
> ?apply
> ?aggregate
```

b. Recreate the data frame from (1a) reporting the number of deaths by country and gender using `aggregate` instead of `dcast`.

```
> data <- read.csv(paste0(main_dir, "data/ebola_fatalities_sex_country.csv"), stringsAsFacto
> aggregate(data$Deaths, list(Country = data$Country, Gender = data$Gender), sum)
Country Gender      x
1       Guinea Female 1001.9
2      Liberia Female 1002.4
3 Sierra Leone Female 3140.0
4       Guinea   Male  930.1
5      Liberia   Male 1055.3
6 Sierra Leone   Male 2987.7
```

c. Keep only rows with data for females and find the total number of deaths across all ages and locations using `apply`.

```
> data <- data[data$Gender == "Female", ]
> apply(data[c("Deaths")], 2, sum)
Deaths
5144.3
```