Project 1

Manav Mandhani, mm58926

Instructions

This knitted Rmarkdown HTML document should be handed in at the beginning of class on **Feb 24th, 2015**. In addition, you *must* hand in the raw Rmarkdown file used to generate the HTML document **via Canvas by 12:30 pm on Feb. 24th**. These two documents will be graded jointly, so they must be consistent (as in, don't change the Rmarkdown file without also updating the knitted HTML!).

All results presented *must* have corresponding code. **Any answers/results given without the corresponding R code that generated the result will be considered absent.** To be clear: if you do calculations by hand instead of using R and then report the results from the calculations, **you will not receive credit** for those calculations. All code reported in your final project document should work properly. Please do not include any extraneous code or code which produces error messages.

For this project, you will be using a recently-compiled dataset reporting fatalities from the on-going Zaire Ebolavirus (Ebola) outbreak in West Africa. You can open the dataset in R directly from the url:

```
ebola <- read.csv("http://wilkelab.org/classes/SDS348/2015_spring_projects/project1/ebola.csv")
head(ebola)
```

```
##                Indicator Country       Date Value
## 1  confirmed_Ebola_cases  Guinea 2014-08-29   482
## 2   probable_Ebola_cases  Guinea 2014-08-29   141
## 3  suspected_Ebola_cases  Guinea 2014-08-29    25
## 4        all_Ebola_cases  Guinea 2014-08-29   648
## 5 confirmed_Ebola_deaths  Guinea 2014-08-29   287
## 6  probable_Ebola_deaths  Guinea 2014-08-29   141
```

```
# Execute the following line of code for R to interpret dates properly. Note that you can now perform summary statistics, etc. with this column. Try it out!
ebola$Date <- as.Date(ebola$Date)
summary(ebola$Date)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2014-08-29" "2014-10-25" "2014-12-10" "2014-11-30" "2015-01-07"
##         Max.
## "2015-02-02"
```

This dataset was compiled by humanitarian workers, mostly based in West Africa, to examine the spread and fatality associated with the current Ebola outbreak over time. This dataset particularly shows the cumulative number of Ebola cases and deaths at certain time periods. Specifically, data was compiled based on whether Ebola was the confirmed, suspected, or probable cause of illness/death (see levels in the `Indicator` column). Note that the levels in this column describing "all" cases and/or deaths indicate the cumulative sum of all confirmed, suspected, and probable Ebola occurrences.

Questions

Question 1: (10 pts) Is this dataset tidy? Explain why or why not. Suggest a different way to represent this data set which *would* be tidy (as in, suggest column names and contents for a data-frame).

No, the data set is not tidy. Column names are not ideal variables and each row is not an object signfying an ebola case. The column names suggested below would make the table tidy.

- Indicator 1
- Indicator 2
- Indicator 3
- Indicator n
- Country
- Date

Question 2: (20 pts) Select **three** (and only three!) indicactor levels from the `Indicator` column in this dataset. Create a new data-frame from the full Ebola dataset such that each of your selected indicators has its own column. The remaining indicators should not be included, but all other values in the dataset should be included. In other words, your final data-frame's columns should be named (with indicators named accordingly) as:

- Indicator 1
- Indicator 2
- Indicator 3
- Country
- Date

Once your data-frame is completed, create a second data-frame from it which has two additional columns displaying summary statistics of your choice. Be sure to display both resulting data-frames using the `head()` function (do *NOT* display the entire data-frame!). You do not need to write a corresponding discussion for this code, but you should incorporate brief explanatory comments throughout your code.

```
ebola %>% filter(Indicator %in% c("confirmed_Ebola_deaths", "probable_Ebola_deaths", "suspected_Ebola_deaths")) %>% spread(Indi
cator, Value) -> ebola1
head(ebola1)
```

```
##   Country       Date confirmed_Ebola_deaths probable_Ebola_deaths
## 1  Guinea 2014-08-29                    287                   141
## 2  Guinea 2014-09-05                    362                   152
## 3  Guinea 2014-09-08                    400                   151
## 4  Guinea 2014-09-12                    403                   150
## 5  Guinea 2014-09-16                    429                   162
## 6  Guinea 2014-09-18                    435                   161
##   suspected_Ebola_deaths
## 1                      2
## 2                      3
## 3                      4
## 4                      4
## 5                      4
## 6                      5
```

```
ebola1 %>% mutate(summary.confirmed = probable_Ebola_deaths/confirmed_Ebola_deaths) ->ebola1.ratio
ebola1 %>% group_by(Country) %>% mutate(summary.mean = mean(confirmed_Ebola_deaths)) -> ebola1.mean
left_join(ebola1.ratio, ebola1.mean) -> ebola1.final
```

```
## Joining by: c("Country", "Date", "confirmed_Ebola_deaths", "probable_Ebola_deaths", "suspected_Ebola_deaths")
```

```
head(ebola1.final)
```

```
##   Country       Date confirmed_Ebola_deaths probable_Ebola_deaths
## 1  Guinea 2014-08-29                    287                   141
## 2  Guinea 2014-09-05                    362                   152
## 3  Guinea 2014-09-08                    400                   151
## 4  Guinea 2014-09-12                    403                   150
## 5  Guinea 2014-09-16                    429                   162
## 6  Guinea 2014-09-18                    435                   161
##   suspected_Ebola_deaths summary.confirmed summary.mean
## 1                      2         0.4912892      924.075
## 2                      3         0.4198895      924.075
## 3                      4         0.3775000      924.075
## 4                      4         0.3722084      924.075
## 5                      4         0.3776224      924.075
## 6                      5         0.3701149      924.075
```

Question 3: (40 pts)

a. (25 points) Use the ggplot2 library to create a plot displaying Ebola cumulative fatality rates over time for **three** countries of your choice (be sure to select countries with at least 4 time-points of data). Fatality rates are computed as the ratio of the number of deaths to the total number of cases. Your plot should **only** consider those values associated with Indicators marked “all_Ebola_cases” and “all_Ebola_deaths”. Your plot should display the points for each country in different colors, and the size of your points should reflect the cumulative number of cases. Your code should be well-commented describing the various steps you take to create this figure.

```
head(ebola)
```

```
##                Indicator Country       Date Value
## 1  confirmed_Ebola_cases  Guinea 2014-08-29   482
## 2   probable_Ebola_cases  Guinea 2014-08-29   141
## 3  suspected_Ebola_cases  Guinea 2014-08-29    25
## 4        all_Ebola_cases  Guinea 2014-08-29   648
## 5 confirmed_Ebola_deaths  Guinea 2014-08-29   287
## 6  probable_Ebola_deaths  Guinea 2014-08-29   141
```

```
summary(ebola$Country)
```

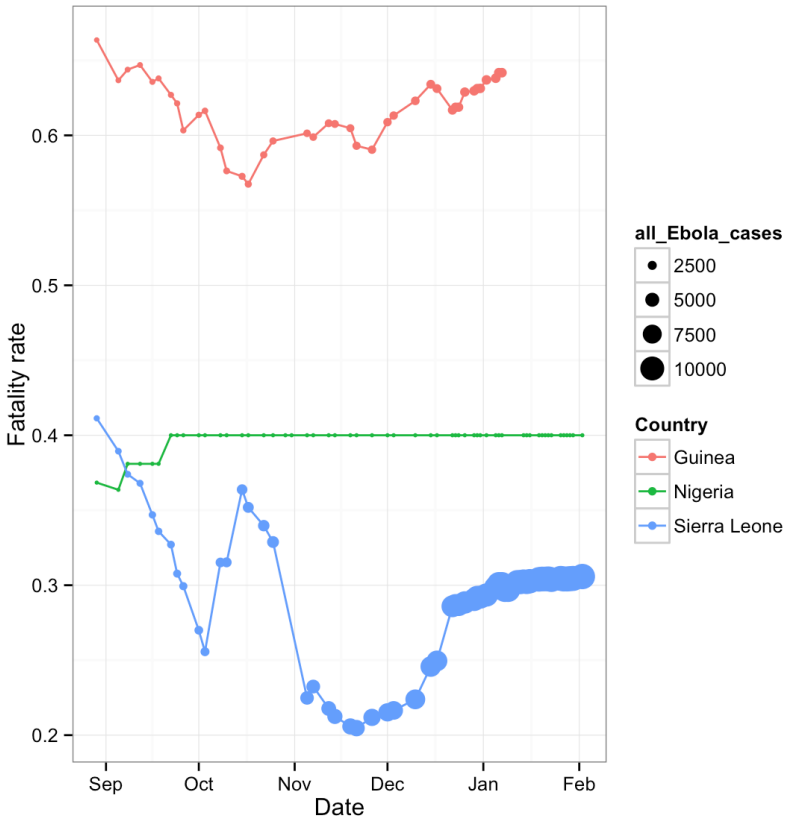
```
##                  Guinea                 Liberia                    Mali
##                     320                     120                      64
##                 Nigeria                 Senegal            Sierra Leone
##                     448                     432                     464
##                   Spain          United_Kingdom United_States_of_America
##                     312                      16                     312
```

```
# Filter out three countries
# Use spread to create indicators as columns
# Select the relevant indicator columns
ebola %>% filter(Country %in% c("Nigeria", "Guinea", "Sierra Leone")) %>% spread(Indicator, Value) %>% select(Country, Date, al
l_Ebola_cases, all_Ebola_deaths)-> ebola_countries

head(ebola_countries)
```

```
##   Country       Date all_Ebola_cases all_Ebola_deaths
## 1  Guinea 2014-08-29             648              430
## 2  Guinea 2014-09-05             812              517
## 3  Guinea 2014-09-08             862              555
## 4  Guinea 2014-09-12             861              557
## 5  Guinea 2014-09-16             936              595
## 6  Guinea 2014-09-18             942              601
```

```
# Calculate fatality rate as y-attribute
ggplot(ebola_countries, aes(x = Date, y = (all_Ebola_deaths)/(all_Ebola_cases), color = Country)) + geom_point(aes(size = all_E
bola_cases)) + ylab("Fatality rate") + geom_line()
```



b. (15 points) Discuss the information (overarching trends, patterns, etc.) your final plot reveals. Be sure to include in your discussion the similarities/differences among countries and a clear, logical justification for why you selected the particular geom(s) used to represent this data. Please limit your full response to a maximum of 6 sentences.

For Sierra Leone, the fatality rate fluctuated massively until a certain threshold toward mid-December where the fatality rate mostly remained stable. It's sudden shifts in fatality rates could be indicative of sudden influx of resources or loss of resources at certain points of time. Nigeria stabilized it's fatality rate in late September and maintained this rate over time as there was no increase in new cases, which can be realized by the fact the size of the points doesn't increase. Guinea suffered a higher fatality rate and seemed to maintain a steady range over time.

I used a scatter plot to plot all the data points in the graph and add a line that connects all the points to indicate trends over time.

Question 4: (30 pts) Think of **two** (and only two!) questions to ask about this Ebola data set. For each question, use the ggplot2 library to create a figure which helps you visualize the trend you're interested in. For each plot, provide a clear explanation as to why this type of plot (e.g. boxplot, barplot, histogram, etc.) is best for representing your chosen trend. Interpret your plot and any trends it reveals, or does not reveal, as the case may be. Your two plots *must* use different primary geoms. Please limit the discussion for each question-plot pair to 4-6 sentences.

Plot One

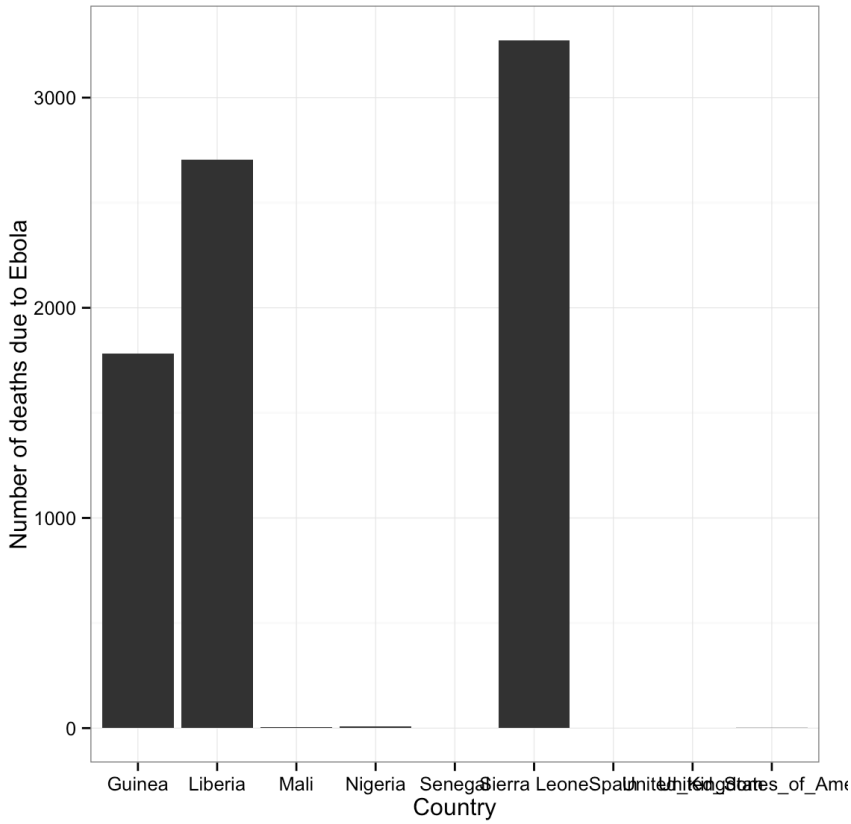
```
head(ebola)
```

```
##                Indicator Country       Date Value
## 1  confirmed_Ebola_cases  Guinea 2014-08-29   482
## 2   probable_Ebola_cases  Guinea 2014-08-29   141
## 3  suspected_Ebola_cases  Guinea 2014-08-29    25
## 4        all_Ebola_cases  Guinea 2014-08-29   648
## 5 confirmed_Ebola_deaths  Guinea 2014-08-29   287
## 6  probable_Ebola_deaths  Guinea 2014-08-29   141
```

```
ebola %>% spread(Indicator, Value) %>% group_by(Country) %>% summarize(sum.country = sum(all_Ebola_deaths)) -> country_frame
head(country_frame)
```

```
## Source: local data frame [6 x 2]
##
##        Country sum.country
## 1       Guinea       45515
## 2      Liberia       25987
## 3         Mali          25
## 4      Nigeria         447
## 5      Senegal           0
## 6 Sierra Leone      116851
```

```
ebola %>% filter(Indicator == "all_Ebola_deaths") %>% group_by(Country) %>% arrange(desc(Date)) %>% slice(1) -> death_frame
ggplot(death_frame, aes(x = Country, weight = Value)) + geom_histogram() + ylab("Number of deaths due to Ebola")
```



I tried to find out which country had the highest number of deaths used to Ebola. The plot above shows that Sierra Leone suffered the highest number of deaths, followed by Liberia and then Guinea.

Plot Two

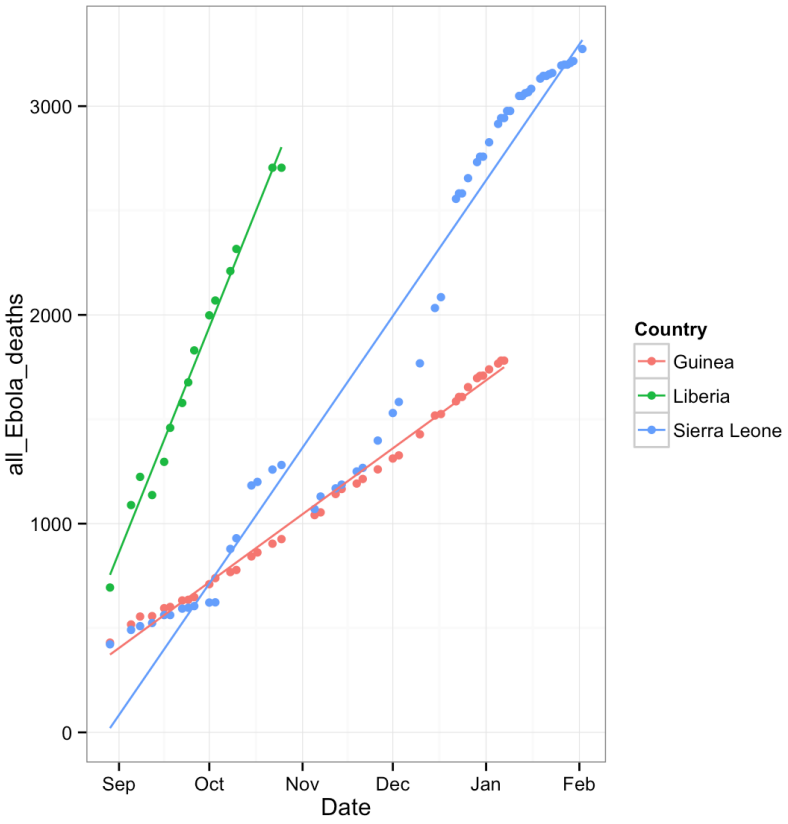
```
head(ebola)
```

```
##                Indicator Country       Date Value
## 1  confirmed_Ebola_cases  Guinea 2014-08-29   482
## 2   probable_Ebola_cases  Guinea 2014-08-29   141
## 3  suspected_Ebola_cases  Guinea 2014-08-29    25
## 4        all_Ebola_cases  Guinea 2014-08-29   648
## 5 confirmed_Ebola_deaths  Guinea 2014-08-29   287
## 6  probable_Ebola_deaths  Guinea 2014-08-29   141
```

```
ebola %>% spread(Indicator, Value) %>% select(Country, Date, all_Ebola_deaths) %>% group_by(Country) %>% filter(all_Ebola_death
s > 20) -> test_ebola
head(test_ebola)
```

```
## Source: local data frame [6 x 3]
## Groups: Country
##
##   Country       Date all_Ebola_deaths
## 1  Guinea 2014-08-29              430
## 2  Guinea 2014-09-05              517
## 3  Guinea 2014-09-08              555
## 4  Guinea 2014-09-12              557
## 5  Guinea 2014-09-16              595
## 6  Guinea 2014-09-18              601
```

```
ggplot(test_ebola, aes(x=Date, y=all_Ebola_deaths, color = Country)) + geom_point() + geom_smooth(aes(group=Country), method=lm
, se=F)
```



My second plot was to find out which country had the highest increase in the number of deaths due to Ebola. The plot was smoothed to try to have a linear view at the data points.

The slope of the charts shows that from September to November, Liberia had the highest increase in deaths, followed by Sierra Leone and then Guinea.