Final

STAT 451-01

Jedidiah Harwood, Kurt Wydrinski

December 11, 2019

Contents

1) Color Pallets

When using color in visualizations there are different kinds of pallets that can be used.

a. Pallet 1 - Sequential Color Schemes with Multi-Hued Colors

Describe the pallet and explain when you would use such a pallet.

Figure 1: Sequential Color Schemes with Multi-Hued Colors

Figure 1 contains examples of *sequential color schemes using multi-hued colors*. Each scheme is sequential because it allows the highlighting of ordering within data through the use of color shading. Within a given hue such as blue, there are multiple shades from light to dark. The lightness or darkness of the hue can be used to represent different levels of values. Typically, light hues represent low values and dark hues represent high values.

Each scheme is multi-hued meaning the colors used are not just shades of a single hue but instead, come from multiple hues. When using multiple hues, the palettes still provide a pleasing aesthetic transition from lighter to darker shades that preserves the implied sequential meaning of the collective colors.

Sequential colors schemes are suited to highlighting data that can be categorized into ordered groups Examples could be age ranges, levels of experience, density ranges, etc. Tile plots are a visualization that typically leverage a sequential color scheme. A key feature of a tile plot (heat map) is display ordinal categorical variables, in a spatial setting. Using a sequential color scheme helps to differentiate catagories providing a well-read, distinguishable, vizualization.

b. Pallet 2 - Diverging Color Schemes

Describe the pallet and explain when you would use such a pallet.

Figure 2: Diverging Color Schemes

Figure 2 contains examples of *diverging color schemes*. Each scheme is diverging because it allows the highlighting of both central and extreme values in underlying data. Lighter shades and hues are used as the central colors in each of these palettes. In the palettes above, each has five (5) colors with the third (3) color being the central, lightest color. Moving away from this color in either direction towards the first and last colors in the palette, the shades and hues get darker. The colors diverge away from a light, neutral

color towards darker, more distinct colors. The colors at the ends of the palettes typically contrast highly from each other to help amplify the meaning of the divergence in the underlying data away from its central values.

Diverging colors schemes are suited to highlight the central and extreme value in data distributions. The light coloring of central values tends to indicate the typical values of data while the bold, contrasting coloring of extreme values tends to highlight these extreme values. Examples could be grade distributions, income level distributions, age ranges, etc.

Note that many datasets could be highlighted by either sequential or diverging schemes. For example, age ranges could be highlighted by either. However, the intent of the visualization would help dictate which to use. Consider a question posed such as, "Comparing pre-teen, teen, adult, and eldery populations. . . ?" Now consider a second, similar question posed such as, "What are the average ages. . . ?" The first question is posed from a categorial perspective that implies a sequence tied to human lifecycles. There is implied interest in order so a sequential color scheme would be applicable. For the second question, there is much emphasis on any difference between young or old but instead, more interest in the distribution, the *average*. In this case, a diverging color scheme may be more suited to the vizualization to not only highlight the average (central values) but also the extreme values.

c. Pallet 3 - Qualitative Color Schemes

Describe the pallet and explain when you would use such a pallet.

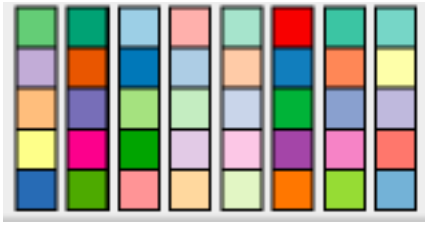


Figure 3: Qualitative Color Schemes

Figure 3 contains examples of *qualitative color schemes*. Each scheme is designed with a set of color shades and hues that contrast from one another. Sequential and diverging color schemes do not try to contrast as much but instead try to show more relationship or transitioning of values between each color. Qualitative schemes try to show the contrast as much as possible attempting to highlight the grouping and differences more than the similarities or nearness to other groups.

Qualitative color schemes are best used when trying to depict different categories of data that are more distinct from each other than they are similar. Examples include demographic data such as racial identity, gender identity, political affiliation, religious affiliation, sports team fan affiliation, etc. These color schemes work well with pie charts, waffle plots, donut plots, and ring plots.

2) Earthquakes

Instructions

Here is the link to the USGS website where the worldwide earthquake data can be downloaded. Download all earthquake data for the past 30 days in .csv format. Using R, make a map of the world with points where the earthquakes occurred. Make a bubble map using the magnitude. Thoroughly discuss your visualizations.

Data Source

The U.S. Geological Survey (USGS) is a program run by the National Institute of Standards and Technology (NIST) to help provide data and information about the occurences of earthquakes. The data is provided in a variety of formats and in a number of frequencies. For this analysis, data on all recorded earthquakes from the 30 days ending November 18, 2019 at 1:08 P.M. PST is being analyzed. This data was obtained from https://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_month.csv at the USGS web site. This data was placed into a file named `earthquake.csv` and is provided along with this report. The table below provides a brief overview of the variables contained with the dataset.

Variable	Description
time	Time of Earthquake occurence
latitude	Latitude Location of Earthquake
longitude	Longitude location of Earthquake
depth	Depth of the Event
mag	Magnitude of Event
magType	Algorithm or Method Used to Evaluate the Method of the Earthquake
nst	Number of Seismic Stations used to evaluate Earthquake Location
gap	The Largest azimutahl gap between azimuthally adjacent stations (in degrees)
horizontalError	Uncertainty of Observed Event's Location (in KM)
dmin	Smallest observed Distance to event epicenter from the Closest Seismic Station
rms	Root Mean Square Calculations of Residuals in predictions of Event occurence.
net	ID of Data Contributor
id	Unique Identification of Eathquake
updated	Time of Upload in Original Dataset
place	Nearby Named Geographical Region
horizontalError	Uncertainty of Earthquake Location (in KM)
depthError	Uncertainty of Earthquake Depth (in KM)
magNst	Total number of Seismic Stations used to Calculate Earthquake's Magnitude
Status	Indicates Whether Event has been viewed by a Person
locationSource	Network that Authored location of Event
magSource	Network that Authored Preferred Magnitude

A small sample of the dataset is provided below. Notice that during this time period of 30 days ending November 18, 2019, there were 11,886 observed earthquakes.

```
## # A tibble: 11,886 x 22
```

```
##    time                latitude longitude  depth   mag magType   nst   gap
##    <dttm>                 <dbl>     <dbl>  <dbl> <dbl> <chr>   <dbl> <dbl>
##  1 2019-11-18 21:01:46     35.8     -119.   8.82  1.14 ml         14    88
##  2 2019-11-18 20:54:46     61.4     -150.  30.4   1.7  ml         NA    NA
##  3 2019-11-18 20:42:44     33.6     -117.  13.4   0.33 ml         17    70
##  4 2019-11-18 20:34:00     33.9     -117.  19.1   0.87 ml         25    67
##  5 2019-11-18 20:32:41     34.4     -118.  11.2   0.83 ml         12   105
##  6 2019-11-18 20:30:32     61.3     -148. 130.    1.6  ml         NA    NA
##  7 2019-11-18 20:27:00     66.3     -157.  16.4   1.7  ml         NA    NA
##  8 2019-11-18 20:11:28     63.5     -147.   0     1.4  ml         NA    NA
##  9 2019-11-18 19:58:04     59.1     -138.   0     2.4  ml         NA    NA
## 10 2019-11-18 19:56:20     35.7     -117.   3.9   0.62 ml         13   144
## # ... with 11,876 more rows, and 14 more variables: dmin <dbl>, rms <dbl>,
## #   net <chr>, id <chr>, updated <dttm>, place <chr>, type <chr>,
## #   horizontalError <dbl>, depthError <dbl>, magError <dbl>, magNst <dbl>,
## #   status <chr>, locationSource <chr>, magSource <chr>
```

Vizualizations

Earthquake Locations

The USGS collects data on earthquakes that occur around the world, not just in the US. Figure 4 shows the locations of the 11,866 observed earthquakes.

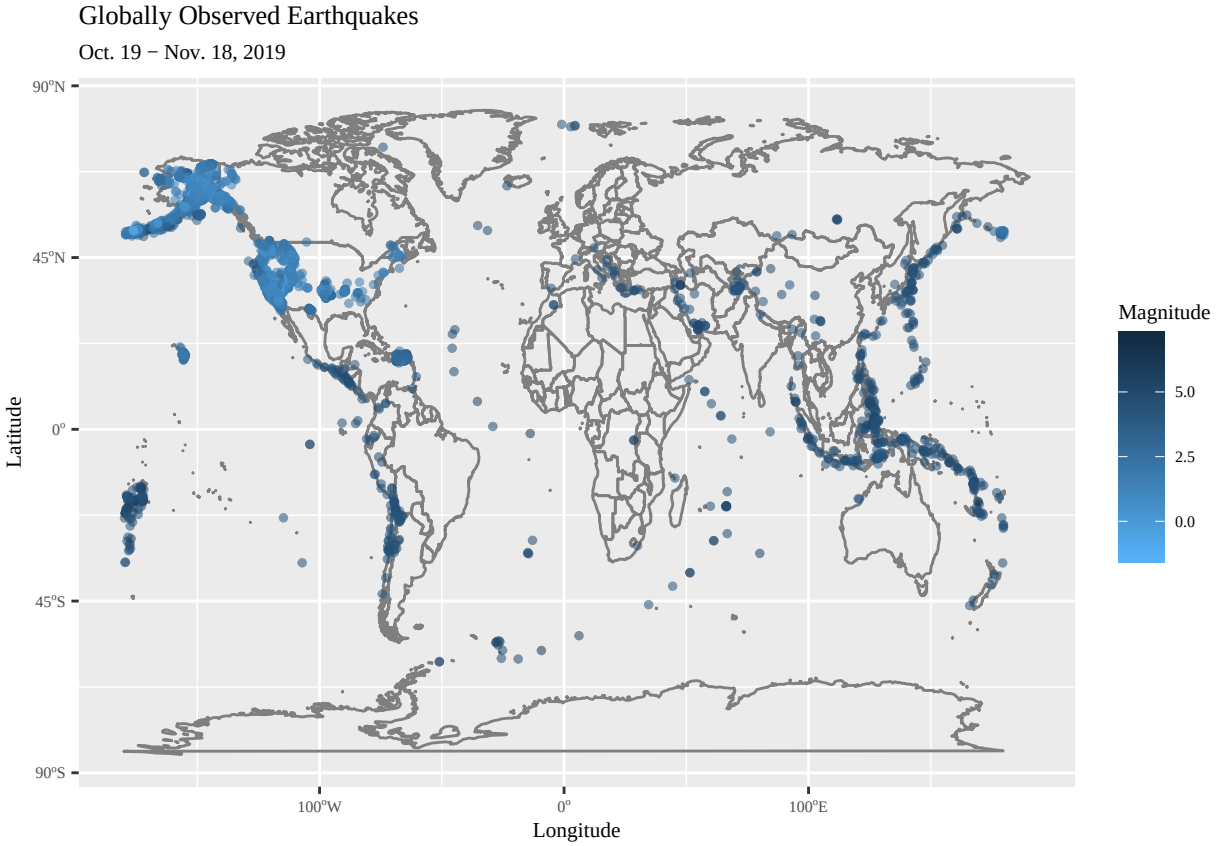


Figure 4: Globally Observed Earthquakes

A cursory view of the map shows clustering earthquakes in typically known locations for earthquakes such as Alaska, the western United States, eastern Mediterranean, and Asian "Ring-of-Fire" around the countries of Japan, Malaysia, Phillipines. A closer inspection reveals that while the United States does have a lot of frequent earthquakes, they tend to be lower in strength (i.e. magnitude) than in other parts of the world. For example, both the western coast of South America and "Ring-of-Fire" show highly concentrated zones of very strong earthquakes.

Earthquake Magnitudes

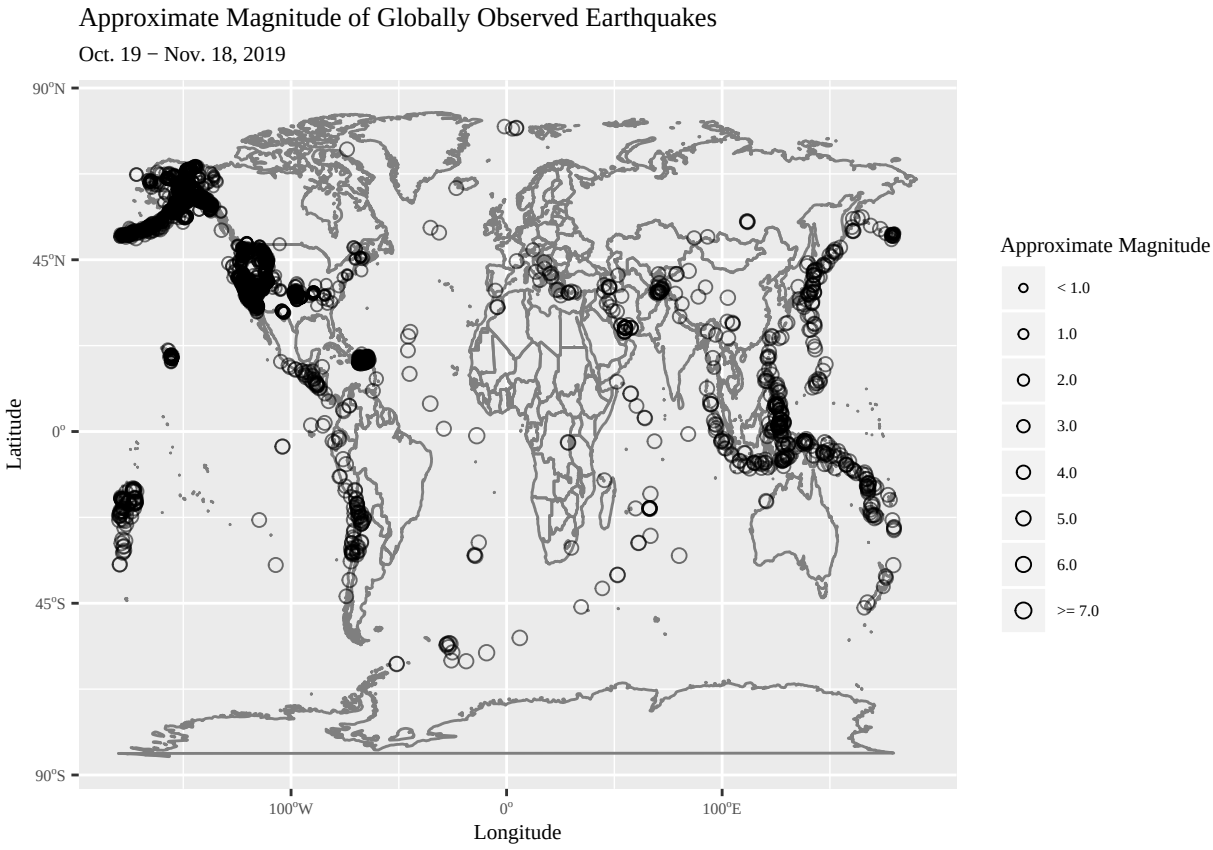


Figure 5: Magnitude of Globally Observed Earthquakes

Figure 5 shows the approximate magnitudes of the observed earthquakes. The size of each circle show the approximate magnitude of each quake.

Figure 6 shows the approximate magnitudes of observed earthquakes in California and nearby areas.

Bubble Map of Earthquakes by Magnitude (With Labels)

```
radius <- sqrt(earthquake$mag / pi)
```

```
## Warning in sqrt(earthquake$mag/pi): NaNs produced
```

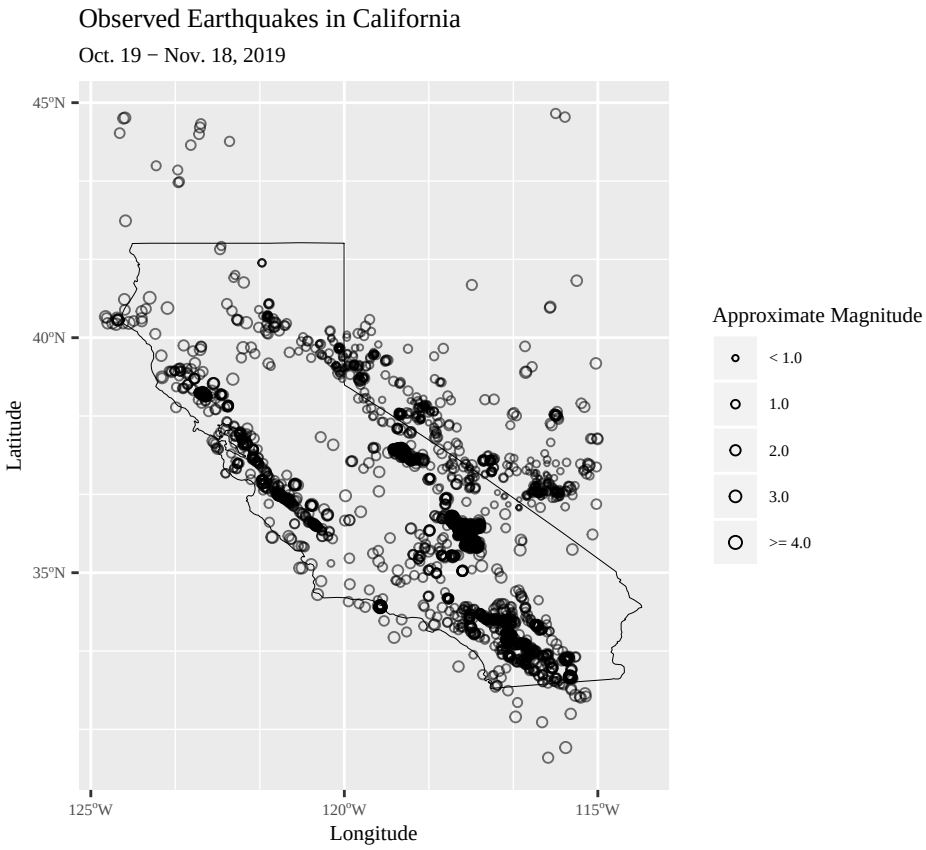
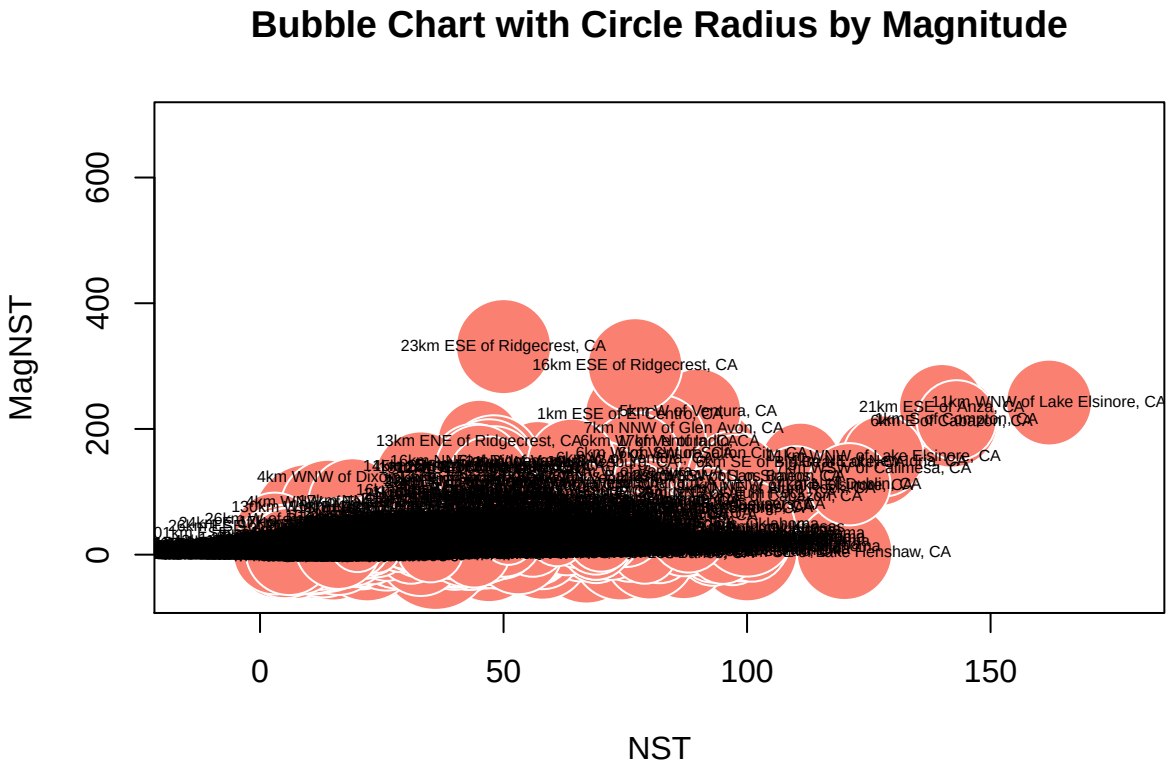


Figure 6: Observed Earthquakes Around California

```r
symbols(earthquake$nst,
        earthquake$magNst,
        circles = radius,
        inches = .35,
        fg = "white", bg = "salmon",
        xlab = "NST", ylab = "MagNST",
        main = "Bubble Chart with Circle Radius by Magnitude")

text(earthquake$nst, earthquake$magNst, earthquake$place, cex = .5)
```



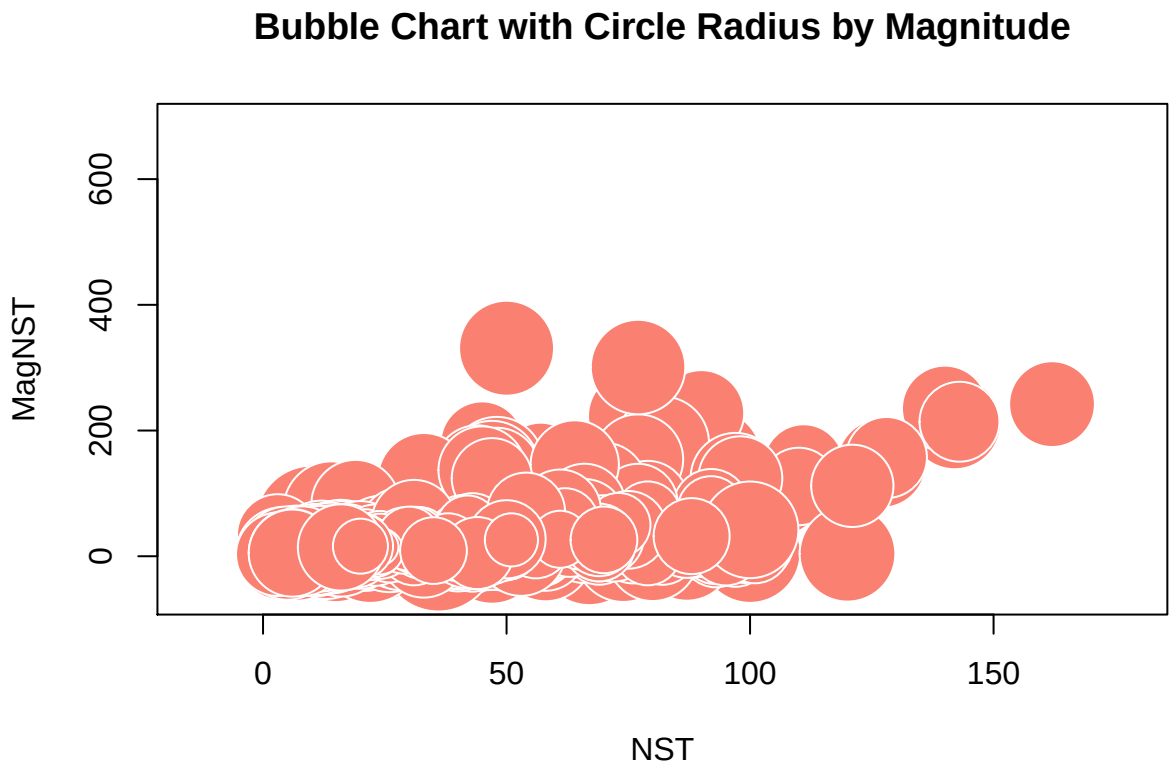
Analysis: For the Bubble Chart, we decided to utilize the variables NST, MAgNST, and the Magnitude. The x-axis represents the amount of seismic sensors used to detect the location, and the y-axis represents the number of seismic sensors used to calculate the magnitude. The diameter of each circle, is based off of the Magnitude for each indivudal observation. As evident from the bubble chart, it appeards that there are more Location seismic sensors used than sensors used to calculate magnitude. In addition, one can tell from the distribuion of the larger circles, that the number of seismic sensors used has no effect on evaluating the magnitude of a quake.

As the clustering of the circle's labels leads to a very distracting plot, the bubble plot has been plotted again, without the text labels.

TODO: KW: I cannot see how magnitude is being plotted. This bubble chart only compares `NST` and `magNST` which are just two type of sensors counts. How should we change this?

Bubble Plot of Earthquakes by Magnitude (Without Labels)

```
symbols(earthquake$nst,
        earthquake$magNst,
        circles = radius, inches = .35,
        fg = "white", bg = "salmon",
        xlab = "NST", ylab = "MagNST",
        main = "Bubble Chart with Circle Radius by Magnitude")
```



TODO: KW: Again, this is teh same as the plot above. How do we document this?

3) Disease / Illness

Data Source

The World Health Organization (WHO) was created shortly after World War II as an international agency whose mission would be to improve overall world health. The WHO works within the United Nations system to help prevent and fight diseases around the world. They maintain a information about this mission in an online database called the *Global Health Observatory (GHO)*. This can be accessed at https://www.who.int/data/gho.

The R package `WHO` provides an interface to the GHO database. This API is can obtain various datasets directly from the database. This analysis will focus on data related to Cholera. Cholera is an infection caused by eating or drinking food or water that is infected the the bacterium *Vibrio cholerae*. While it is preventable and treatable, it can cause death. The WHO estimates there are upward of 4 million cases of the infection with upwards of 143,000 of these resulting in death. See https://www.who.int/health-topics/cholera#tab=tab_1 for further details.

The R API was used to collect data related to cholera. Observations for the following indicators were collected and analyzed.

Indicator	Description	Further Details
CHOLERA_0000000001	Number of reported cases of cholera	https://www.who.int/data/gho/indicator-metadata-registry/imr-details/42
WSH_10	Number of diarrhoea deaths from inadequate water, sanitation and hygiene	https://www.who.int/data/gho/indicator-metadata-registry/imr-details/2260

Vizualizations

Cases of Cholera by Deaths from Improper Water

The smoothed scatterplot in figure 7 depicts that as the number of cases of Cholera increases, the number of deaths from a basic water source remains relatively constant. This may be counter-intuitive from certain perspectives that may think that there should be a strong, positive linear correlation between Cholera cases and related deaths. However, it appears that the amount of Cholera cases is at a constant trend with the amount of deaths in a country.

Reported Cholera Cases by Country

The Lollipop chart in figure 8 depicts that that the country with the most reported Cholera cases, is Haiti. Other countries with significant Cholera populations, include the Democratic Republic of the Congo, Yemen, Somalia, the United Republoic of Tanzania, Kenya and South Sudan. For the most part, it appears that Cholera is not very prevalent among other countries.

Water Quality Related Deaths

Figure 9 presents a histogram of deaths related to poor water quality by country. The graph also uses a continuous color scale to repesent the number of reported Cholera cases within each country. Dark colors on the scale indicate more reported cases than lighter colors.

It appears that the countries with the most water-quality related deaths, are India, Nigeria, and the Democratic Republic of the Congo. It also appears that countries with the most water quality related deaths do not differ from countries with the least amount of water-quality related deaths, with respect to their rported

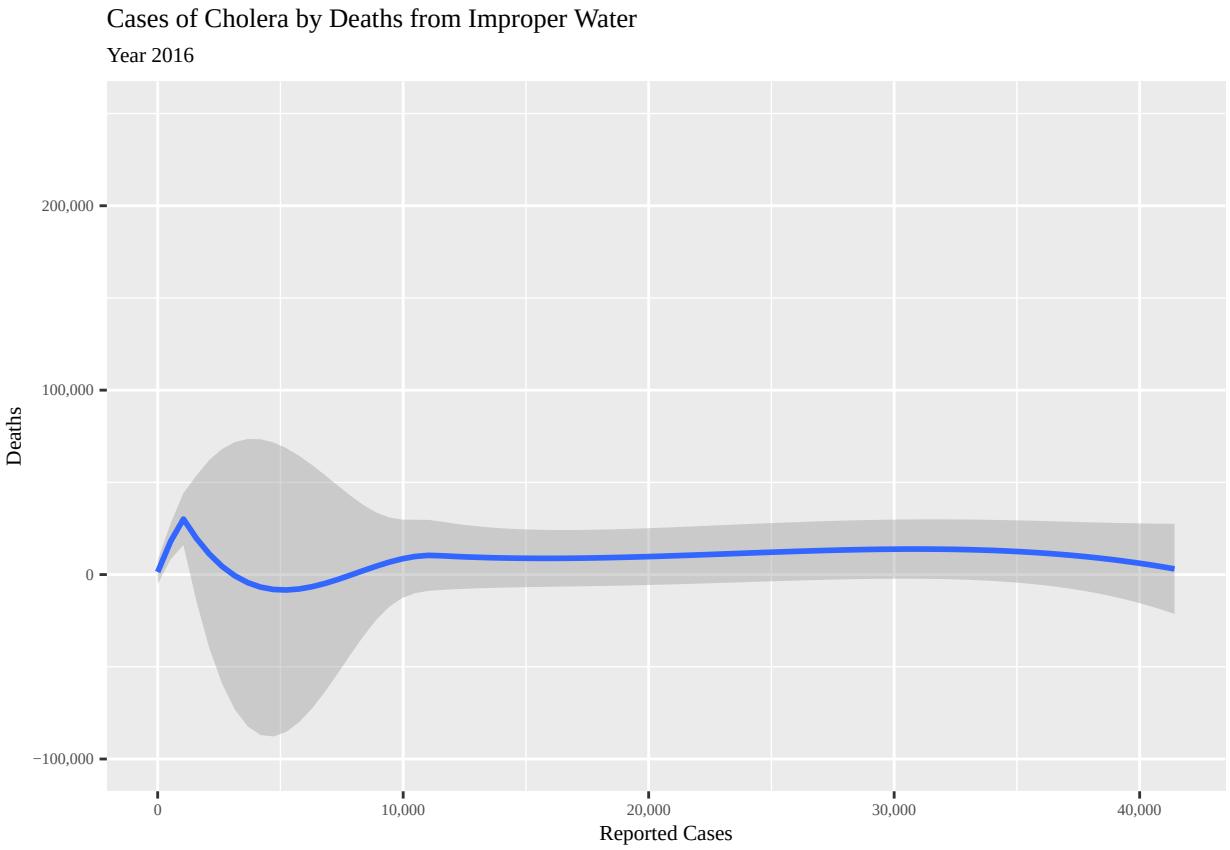


Figure 7: Cases of Cholera by Deaths from Improper Water

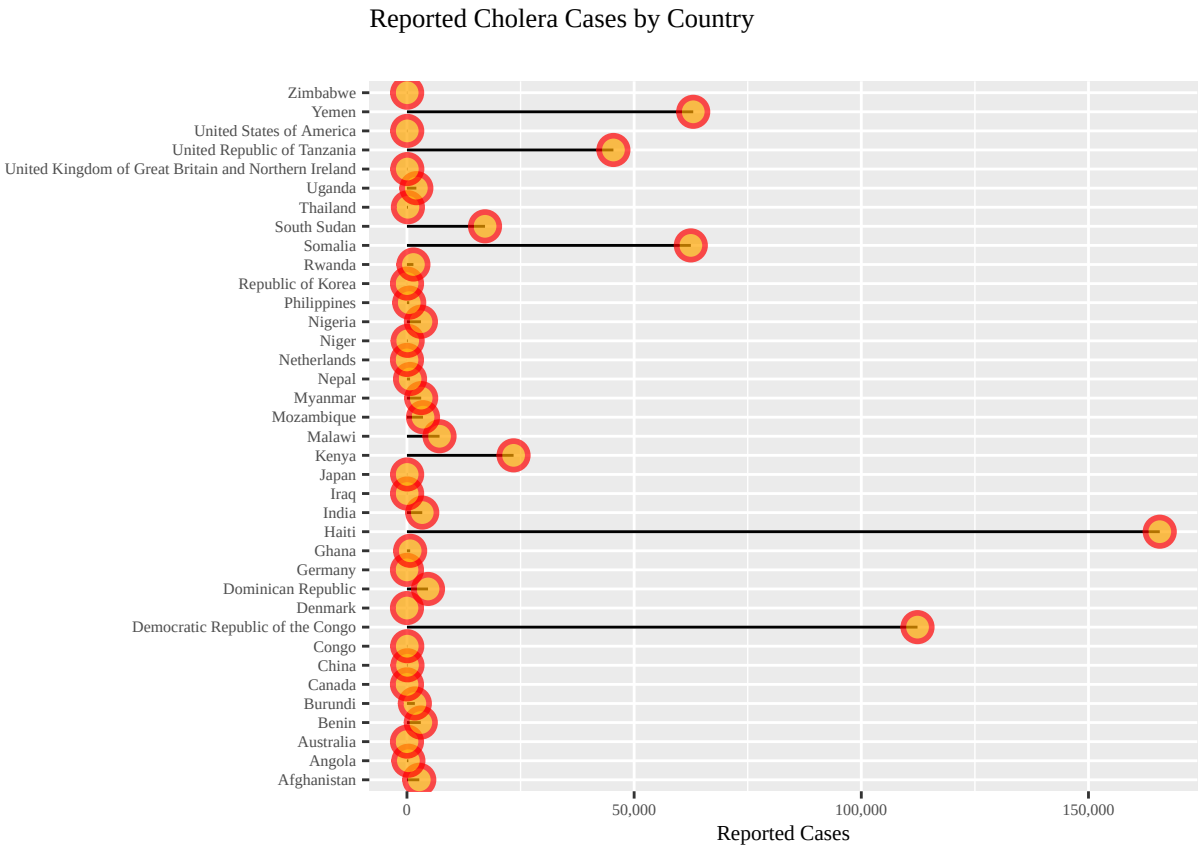


Figure 8: Reported Cholera Cases by Country

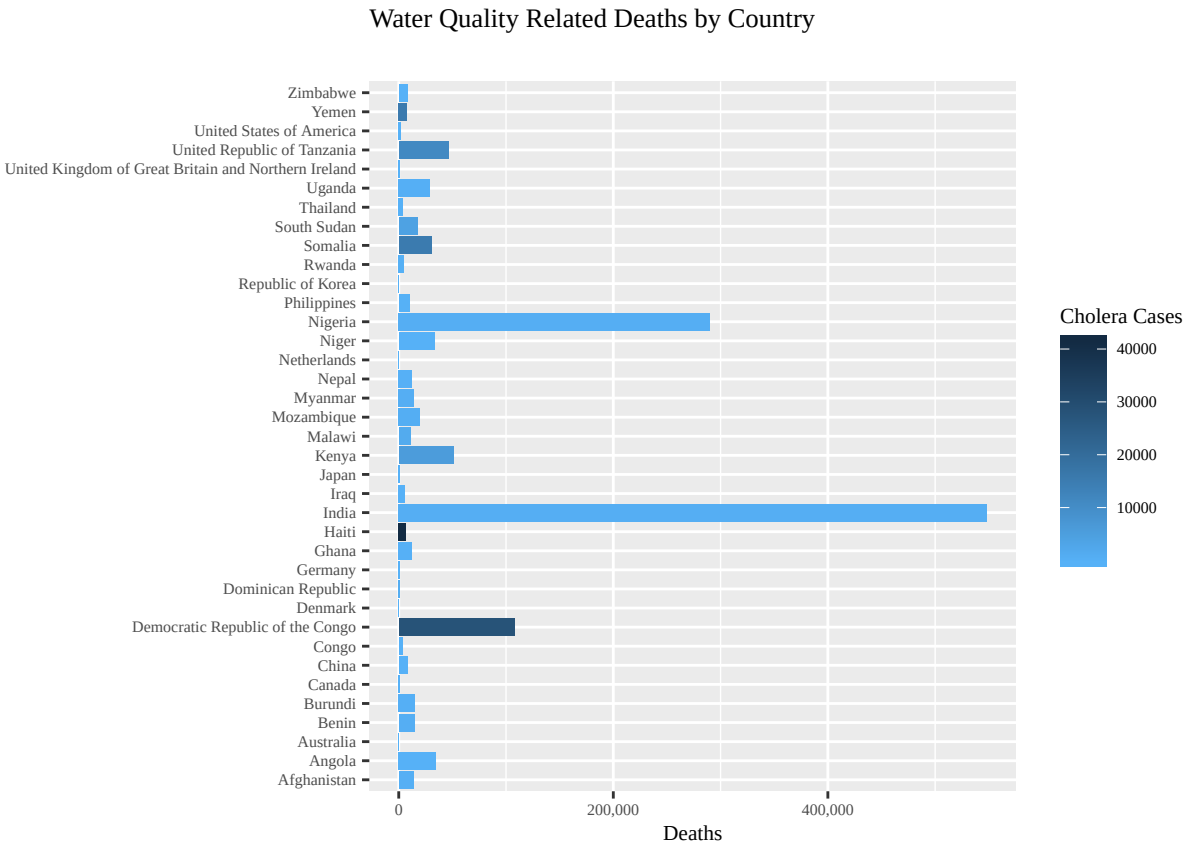


Figure 9: Water Quality Related Deaths by Country

numbers of cholera cases. However, one exception is Kenya. From this observation, one may infer that the number of water-quality related deaths may not be as correlated to the number of Cholera cases, as expected. It is likely that other factors such as the availability of water treatment facilities may influence the number of deaths caused by poor water quality within each country.

Regional Share of Cholera Cases

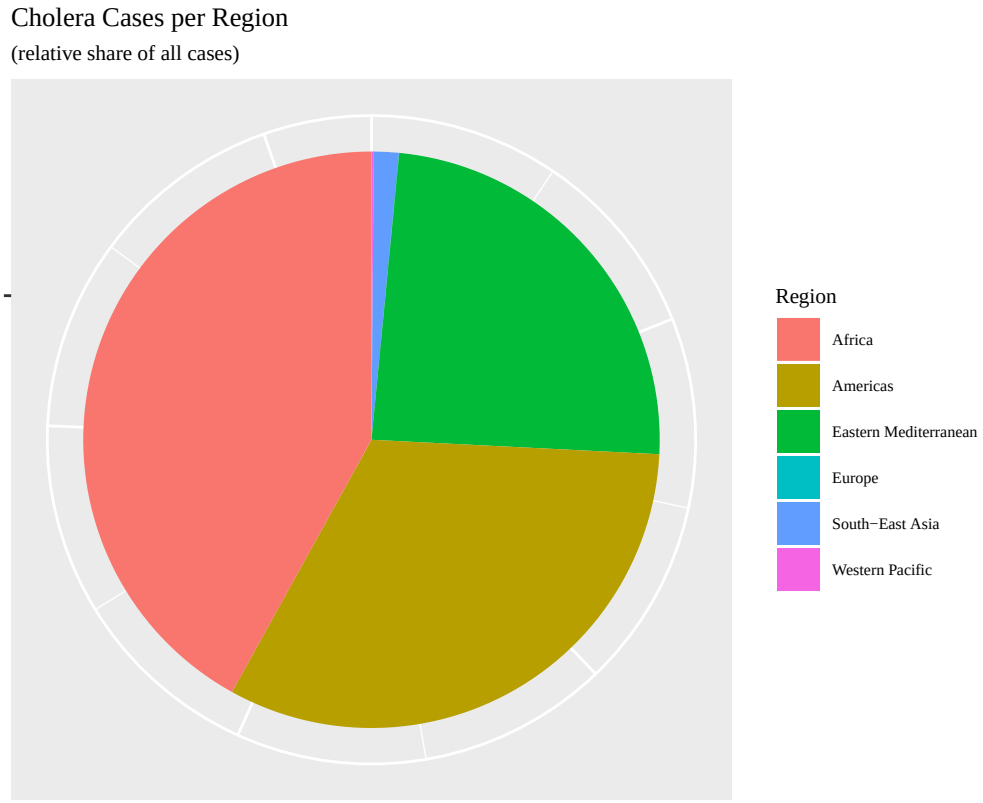


Figure 10: Regional Share of Cholera Cases

The pie chart in figure 10 shows the relative share of reported cholera cases for each region. Africa has the largest share of cases followed closely by the America then the Eastern Mediterranean. The region with the least amount of observed Cholera cases, was Europe. The sliver for Europe is so small, it is not visible on this pie chart.

Data Wrangling

Although one would think the data maintained by the WHO in the GHO database is clean and possibly even tidy, it is not. This section describes how this data was retrieved, transformed and prepared for the preceding analysis and visualizations.

First, the `CHOLERA_0000000001` dataset containing the observed number of reported cases of Cholera is read from the GHO database.

```r
tb_cholera <- get_data("CHOLERA_0000000001")

tb_cholera <- tb_cholera %>%
  group_by(country) %>%
  arrange(country, year) %>%
  select(country, year, value, region) %>%
  rename("cases" = value)
```

Next, the `WSH_10` dataset containing the observed deaths due to poor water quality was then downloaded. This data is joined to the reported cholera cases data. However, it is important to note that the WSH-10 dataset only contains observations for 2016.

```r
d_table <- get_data("WSH_10")

viz_data <- tb_cholera %>%
  left_join(d_table,
            by = c("country", "year", "region")) %>%
  filter( year == 2016) %>%
  rename('deaths' = value)
```

Although the data appears tidy and is close to being usable, the `deaths` variable contains extra data besides the death counts that needs to be ignored. The death counts are parsed out of the data to make the data yet even closer to being ready to analyze.

```r
viz_data$deaths <- viz_data$deaths %>%
  str_remove(pattern = "[:space:]\\[[:digit:]+-[:digit:]+\\]") %>%
  parse_integer()
```

There is one variable `gho` that exists in the original dataset but it not needed. This variable is removed leaving the `viz_data` dataset as tidy and ready for the above analysis.

```r
viz_data <- viz_data %>% select(-gho)
```