Data Analysis

Chen Chen

11/29/2017

After loading the earthquake data, the population data, and the GDP data into R studio, I plotted the following plots:

Correlation of the earthquake data

Total deaths and deaths are strongly correlated. The death count and the death description are strongly correlated as well.

Distrubtion of total deaths

This is a J-shaped distribution. The Sumatra earthquake in 2004 had the highest total death of 227899, and the M 7.6 earthquake in Pakistan in 2005 had the second hightest death toll of 76213.
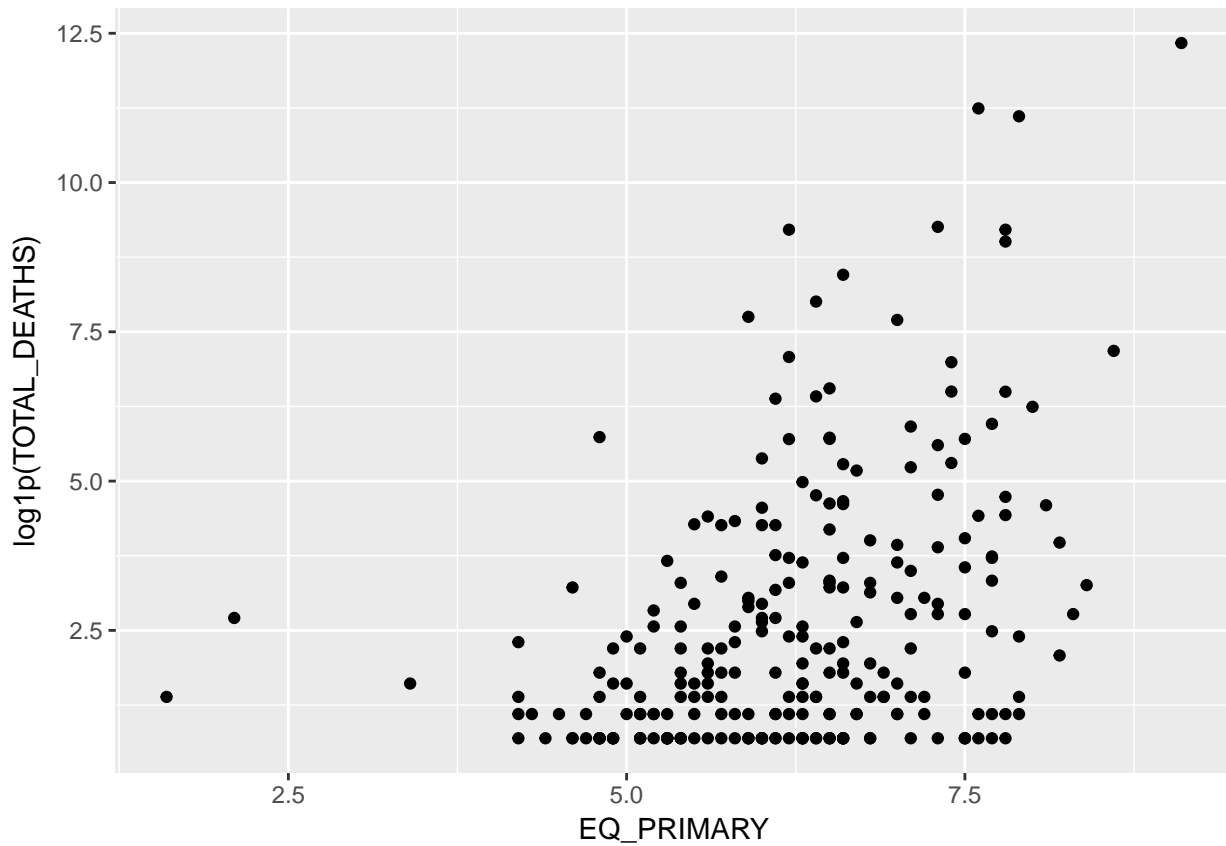
```
ggplot(df_eq_gdp,aes(x=log1p(TOTAL_DEATHS))) +
  geom_histogram(bins = 50,color="black",fill = "white")
```
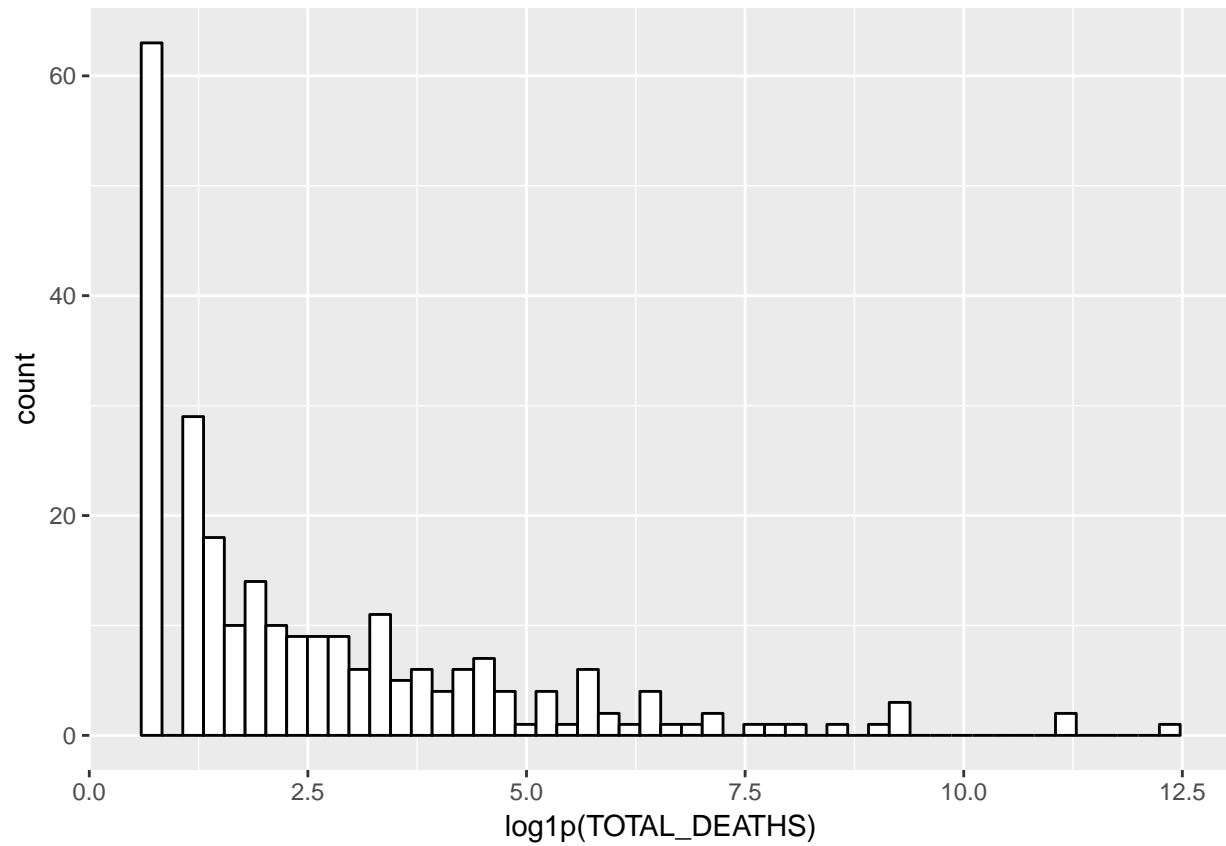
Total deaths VS Earthquake Magnitude

There is a positive correlation of earthquake magnitude and the total death in an earthquake. As expected, more people died in large earthquakes.

```
ggplot(country_eq,aes(x=EQ_PRIMARY, y = log1p(TOTAL_DEATHS))) +
  geom_point()
```
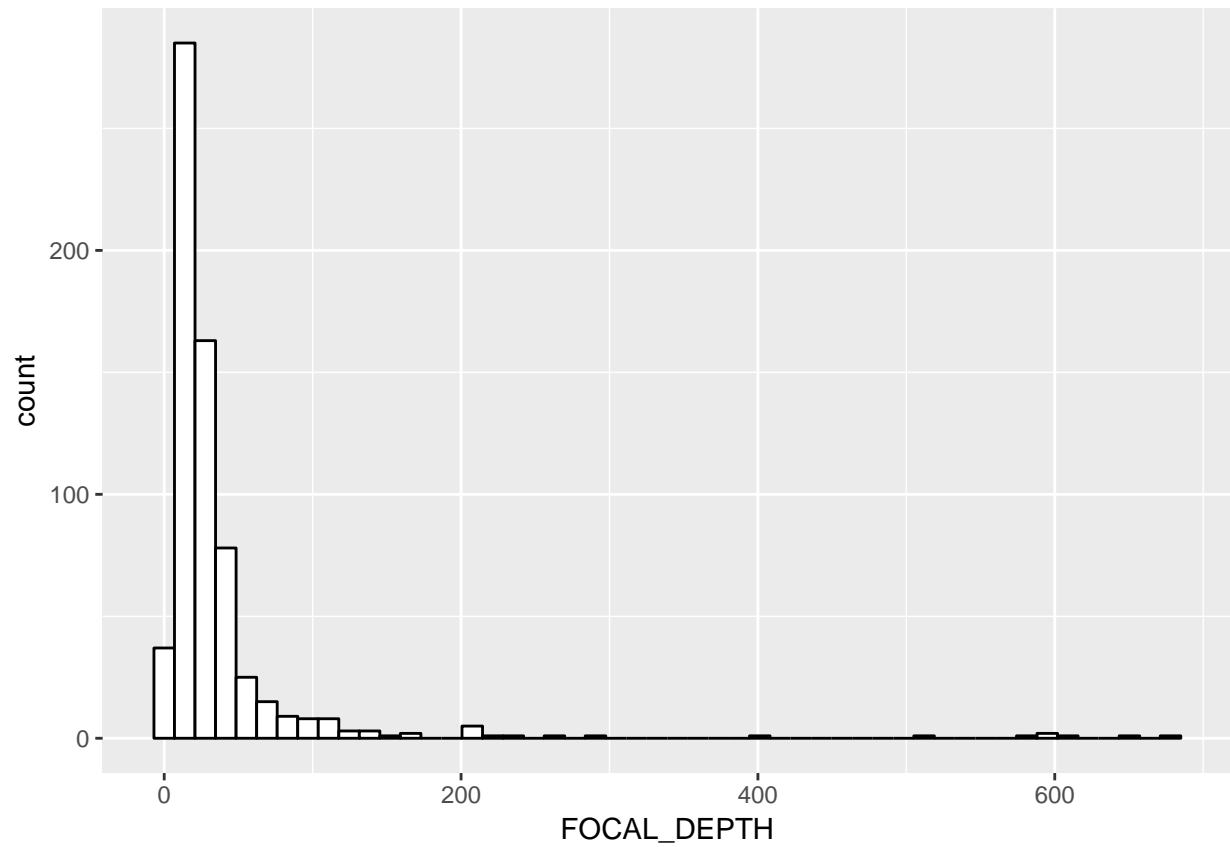
Total deaths VS Focal Depth of earthquakes

There is a weak correlation between focal depth and total deaths. More deaths are correlated with shallower earthquakes. There is usually no death in very deep earthquakes (> 300 km), except for one earthquake in Peru that occurred in 1970 had 1 death on record.
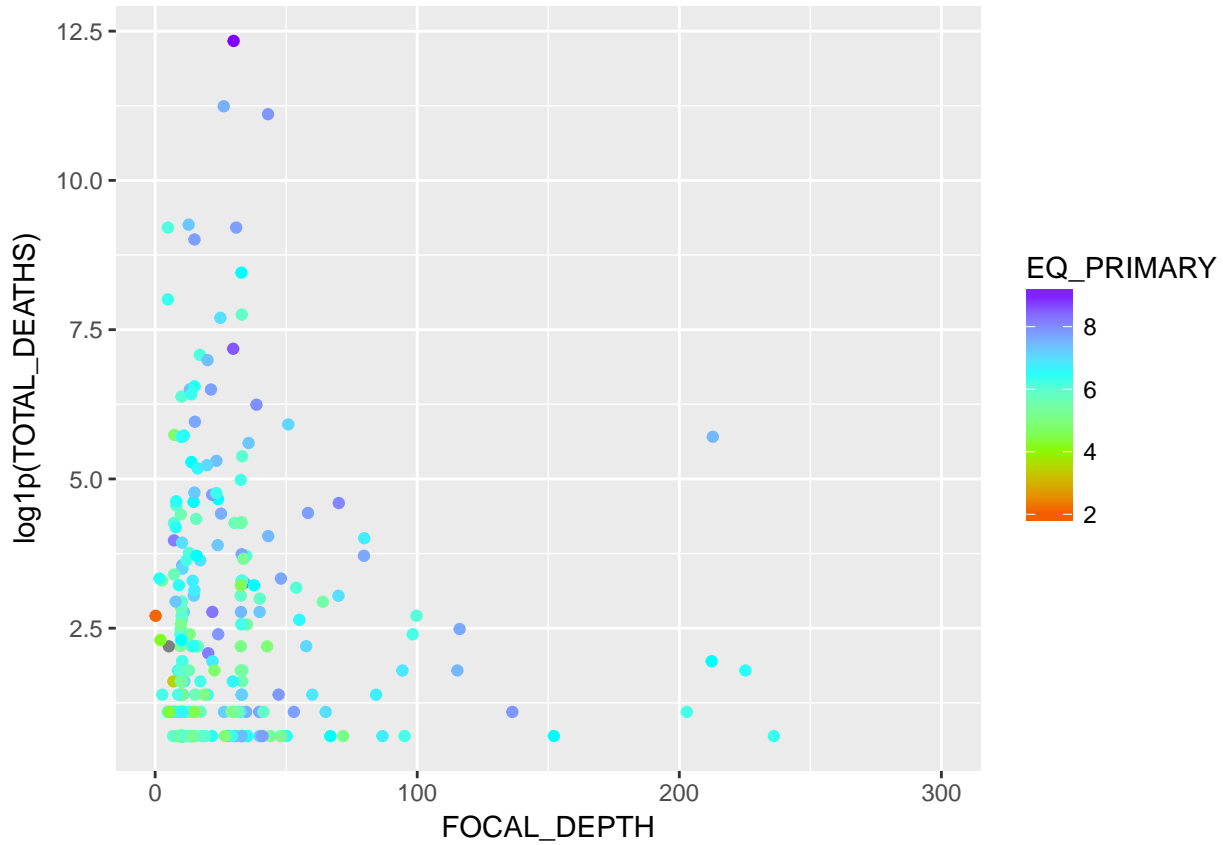

It will be useful to compare this result with the distribution of earthquake focal depth, as shown below.

Because in this study we are investigating significant earthquakes, i.e. earthquakes that have caused damages, the distribution of earthquake focal depth is right skewed, with a few deep earthquakes that have had deaths.

Total damage VS Focal Depth of earthquakes

There is no obvious correlation between focal depth and total damages. But it is interesting to see that only earthquakes that are shallow than 100 km depths have caused damages.
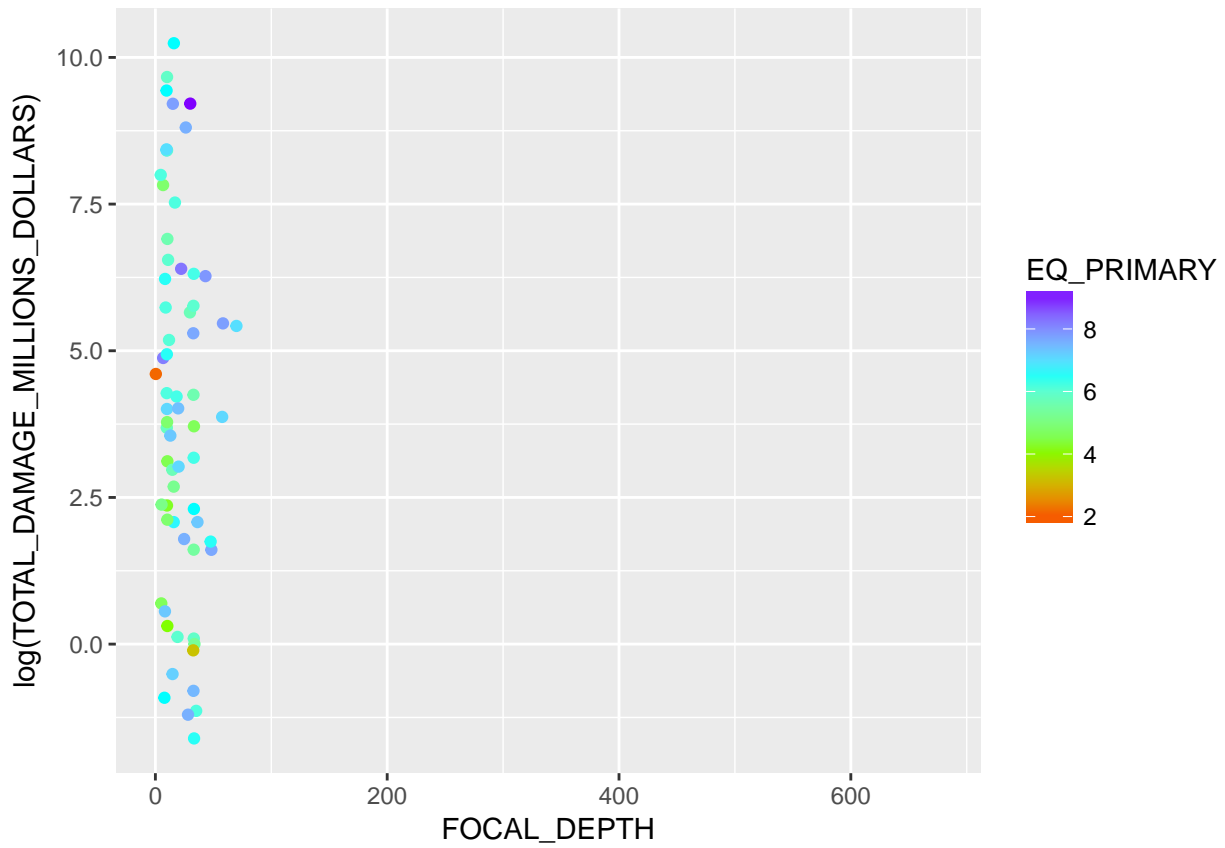
Total Damage VS Total Deaths

There is a weak correlation of higher damage with higher death rate.

```
ggplot(country_eq,aes(x=log(TOTAL_DAMAGE_MILLIONS_DOLLARS),y=log(TOTAL_DEATHS))) +
  geom_point()
```
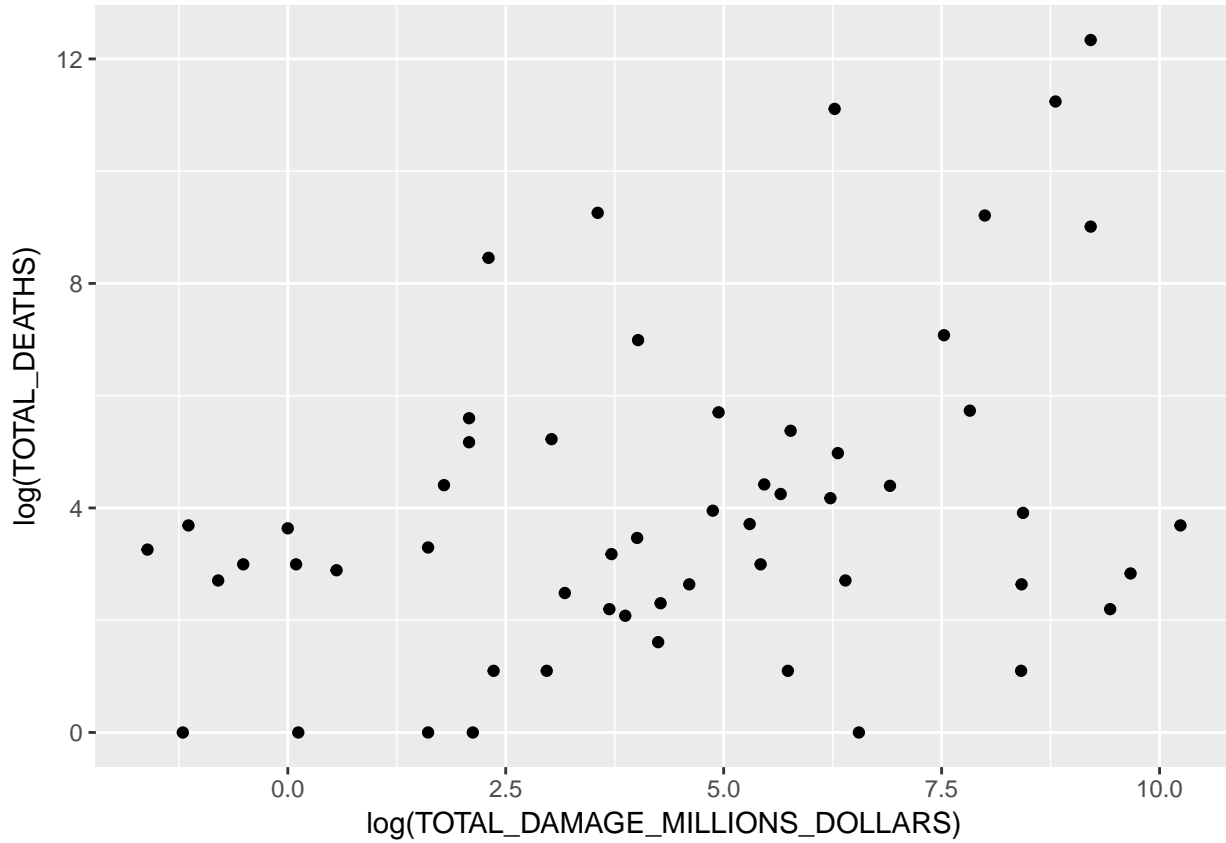
Total deaths over the years

There is no decrease of total deaths in large earthquakes over the years. The earthquake data seems to be missing the 1980s data. Inspecting the data frame, and there seems to be a hole in the data. Inspecting the database revealed that the error occurred during the step of data import. Will need to fix this.

```
ggplot(country_eq, aes(x= format(DATE,"%Y"), y = log(TOTAL_DEATHS))) +
  stat_summary(fun.y=sum,geom="bar") +
  theme(axis.text.x = element_text(angle = 90)) +
  scale_x_discrete("Year")
```
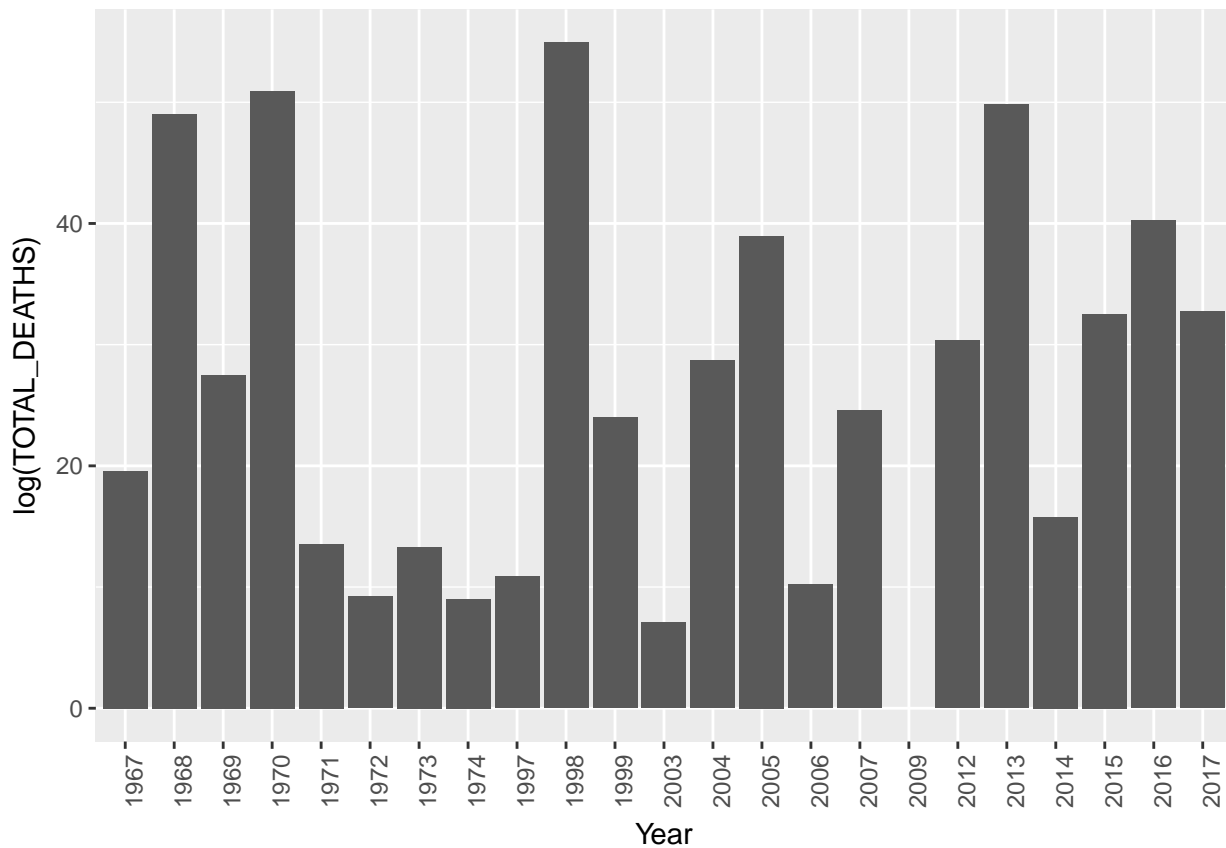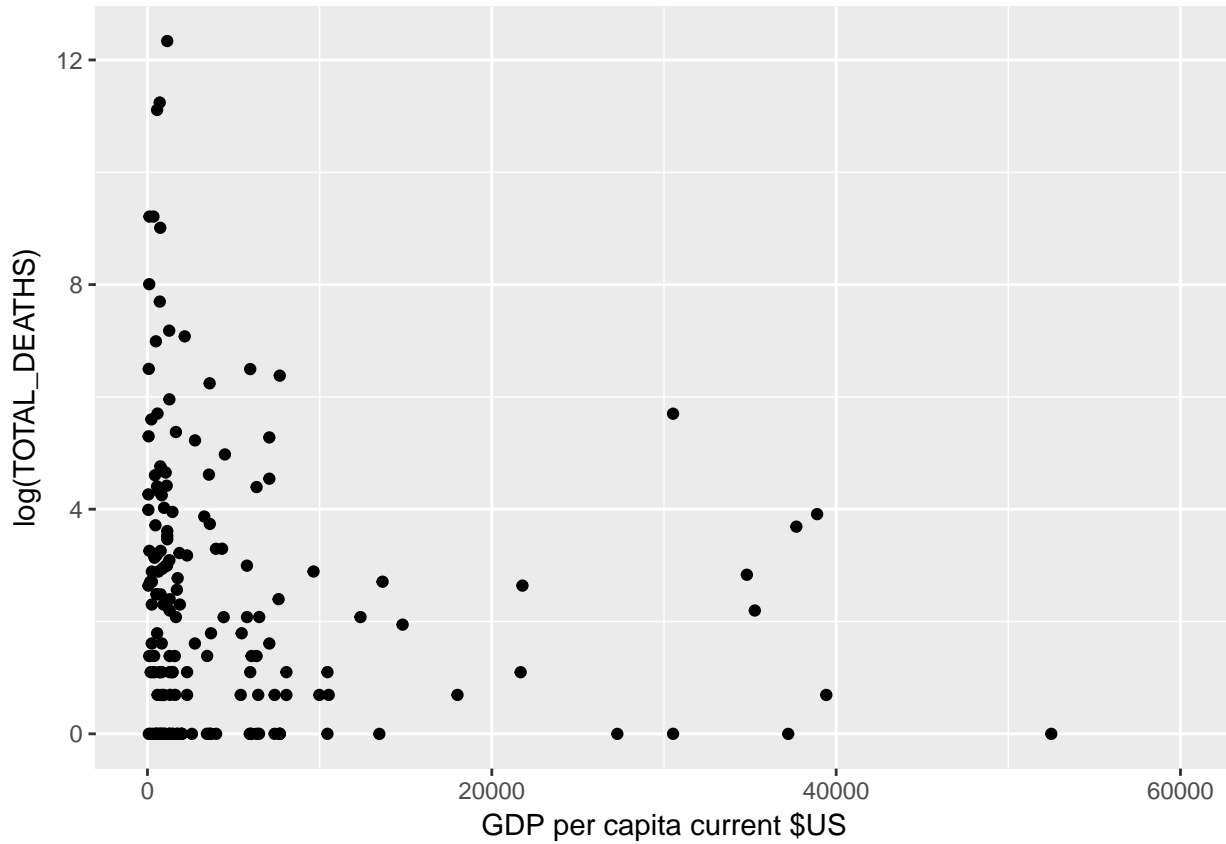
Total deaths VS GDP per capita

The data seems to show that higher death rates are associated with lower GDP per capita.

```
ggplot(df_eq_gdp,aes(x = as.numeric(GDP_per_capita_currentUSD),y=log(TOTAL_DEATHS))) +
  geom_jitter() +
  scale_x_continuous("GDP per capita current $US",limits = c(0,60000))
```

Total deaths VS Country population

```r
ggplot(df_eq_gdp,aes(x = as.numeric(Population),y=log(TOTAL_DEATHS))) +
  geom_jitter() +
  scale_x_continuous("Country Population")
```