Assignment 2: Visualization for Numeric Data

Maya Maciel-Seidman

2024-02-22

Set Up

```
# Load tidyverse:
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#Load knitr:
library(knitr)

#Load dplyr:
library(dplyr)

#Load ggplot2:
library(ggplot2)

#Load ggpubr:
library(ggpubr)
```

Data and Data Wrangling

```
earthquakes <- read.csv("./earthquakes.csv")
```

This dataset is NOAA's significant Earthquake Database from 2150 BC to October 16, 2017. It was downloaded as a csv from benjiao's GitHub page at this link: https://github.com/benjiao/significant-earthquakes/blob/master/earthquakes.csv. This dataset records all significant earthquakes around the globe

that occurred in the timeline described above. For each earthquake, the dataset contains characteristics including magnitude, focal depth (depth of the earthquake's hypocenter), death toll, and location, among other variables. The dimensions of this dataset are 5963 x 16. This means that there are 5,963 observations (earthquakes) and 16 variables describing each earthquake.

Research Question

Using this dataset, I aim to investigate if an earthquake's focal depth (depth of hypocenter below Earth's surface) or magnitude is more strongly correlated with the number of deaths resulting from that earthquake. Additionally, I aim to determine the distribution of the death toll between earthquake magnitude classes (categorical descriptions of an earthquake's magnitude) to determine if death toll follows the severity of earthquake magnitude.
Null hypothesis 1: An earthquake's magnitude is just as correlated to death toll as its focal depth is.
Alternative hypothesis 1: One variable (magnitude or focal depth) will be more strongly correlated to an earthquake's death toll.
Null hypothesis 2: The distribution of deaths between magnitude classes does not follow an earthquake's severity.
Alternative hypothesis 2: The distribution of deaths between magnitude classes does follow an earthquake's severity.

Variables of Interest

The variables of interest to address my research question are an earthquake's focal depth, magnitude, and death toll. The focal depth is the depth at which an earthquake occurs and its data type is an integer in this dataset. The magnitude is a measure of an earthquake's strength and its data type is a double in this dataset. Finally, the death toll is how many deaths occurred as a result of an earthquake and its data type is also an integer.

Data Wrangling

```
# Select only the columns with the variables of interest:
earthquakes <- earthquakes %>% select(deaths, focal_depth, magnitude)

# Look at summary of the data to find if there are NA values:
summary(earthquakes)
```

```
##      deaths        focal_depth       magnitude
##  Min.   :     1   Min.   :  0.00   Min.   :1.600
##  1st Qu.:     4   1st Qu.: 11.00   1st Qu.:5.700
##  Median :    26   Median : 27.00   Median :6.500
##  Mean   :  3988   Mean   : 41.94   Mean   :6.489
##  3rd Qu.:   400   3rd Qu.: 40.00   3rd Qu.:7.300
##  Max.   :830000   Max.   :678.00   Max.   :9.500
##  NA's   :4019     NA's   :2945     NA's   :1789
```

```
# Get rid of NA values:
earthquakes <- earthquakes %>% drop_na()
```

```r
# Ensure that there are no more NA values:
summary(earthquakes)
```

```
##      deaths          focal_depth       magnitude
##  Min.   :     1.0   Min.   :  0.00   Min.   :1.600
##  1st Qu.:     2.0   1st Qu.: 10.00   1st Qu.:5.700
##  Median :    10.0   Median : 24.00   Median :6.400
##  Mean   :  2257.8   Mean   : 32.45   Mean   :6.404
##  3rd Qu.:    82.5   3rd Qu.: 34.00   3rd Qu.:7.200
##  Max.   :316000.0   Max.   :651.00   Max.   :9.500
```

```r
# Create a new column for magnitude class since the original dataset did not
# have a column for this:
earthquakes <- earthquakes %>%
  mutate(magnitude_class =
           case_when(magnitude < 3 ~ "Micro",
                     magnitude >= 3 & magnitude <= 3.9 ~ "Minor",
                     magnitude >= 4 & magnitude <= 4.9 ~ "Light",
                     magnitude >= 5 & magnitude <= 5.9 ~ "Moderate",
                     magnitude >= 6 & magnitude <= 6.9 ~ "Strong",
                     magnitude >= 7 & magnitude <= 7.9 ~ "Major",
                     magnitude >= 8 ~ "Great"))

# Check the first few rows of the data to make sure the new column was added properly:
head(earthquakes)
```

```
##   deaths focal_depth magnitude magnitude_class
## 1      1          18       7.1           Major
## 2  12000          10       5.2        Moderate
## 3  12000          10       5.2        Moderate
## 4  50000          15       7.6           Major
## 5  60000          10       6.5          Strong
## 6   2000          10       6.1          Strong
```

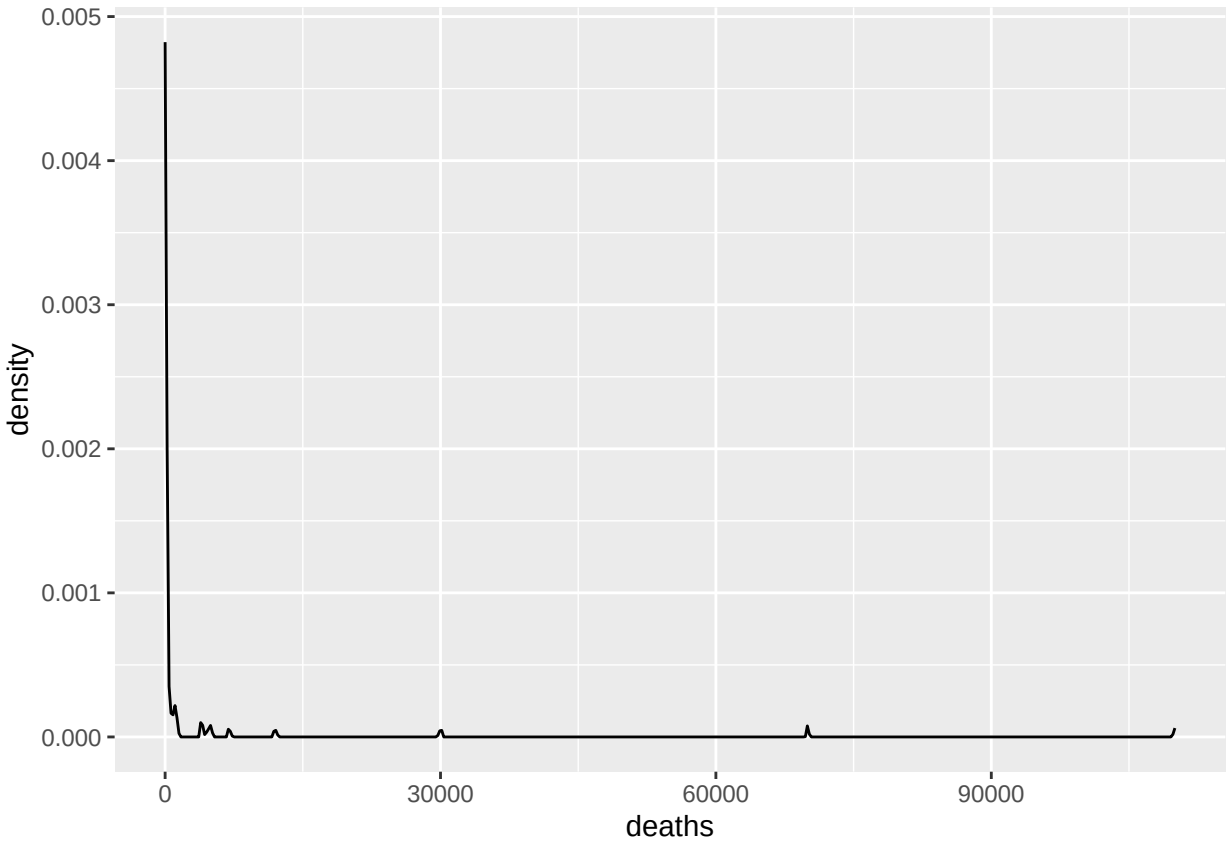
```r
# Set seed for consistent random sample:
set.seed(123)
# Take a random sample of 100 earthquakes:
earthquakes_sample = sample_n(earthquakes, 100)
```

First, I used `select()` to create a subset of the original dataset with only the columns containing variables of interest. I used `summary()` to see a statisitcal summary of the new dataset and found that there were NA values. I then removed all NA values using `drop_na()` and double checked to make sure no NA values were left by using `summary()` again. Then, I created a new column in the dataset for the magnitude class (`magnitude_class`), which is a categorical variable that describes the magnitude of an earthquake in a way that gives more context than just the numerical Richter Scale value of magnitude. Magnitude classes are based on ranges within the Richter Scale magnitude values. I created this new column using `mutate()` and `case_when()`, which allowed for conditional statements to determine the values of the new column. Then, I used `head()` to check the first few rows of the dataset to ensure that the new column was created correctly. Finally, I took a random sample of 100 earthquakes since the full dataset would have produced messy graphs with too many datapoints than necessary to view trends in the data.

Visualization

Preliminary Check for Skew in Death Toll

```
# Create density plot for death toll:
ggplot(data=earthquakes_sample, aes(x=deaths)) + geom_density()
```



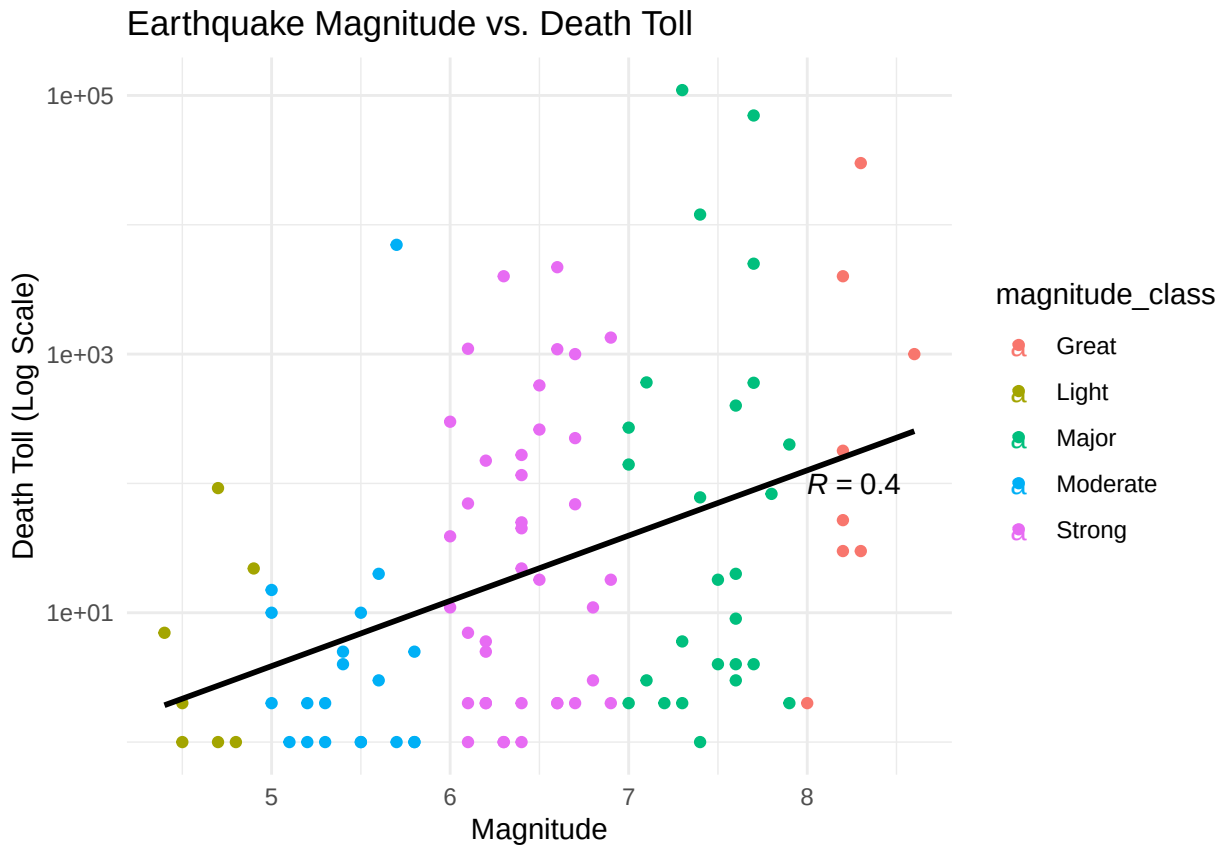

```
# Death toll has a strong right skew
```

Scatter Plot for Magnitude and Death Toll

```
# Create scatter plot for magnitude vs. deaths:
ggplot(data=earthquakes_sample, aes(x=magnitude, y=deaths, color=magnitude_class, group=1)) + # color t
  geom_point() + # scatter plot
  scale_y_log10() + # log base 10 the deaths variable due to skew
  geom_smooth(method=lm, color="black", se=FALSE) + # linear regression line with no standard error
  labs(title="Earthquake Magnitude vs. Death Toll", x="Magnitude", y="Death Toll (Log Scale)") + # add
  theme_minimal() + # add minimalist theme
  stat_cor(aes(label = ..r.label..), label.x = 8, label.y=2) # display Pearson correlation coefficient
```

```
## Warning: The dot-dot notation (‘..r.label..‘) was deprecated in ggplot2 3.4.0.
```

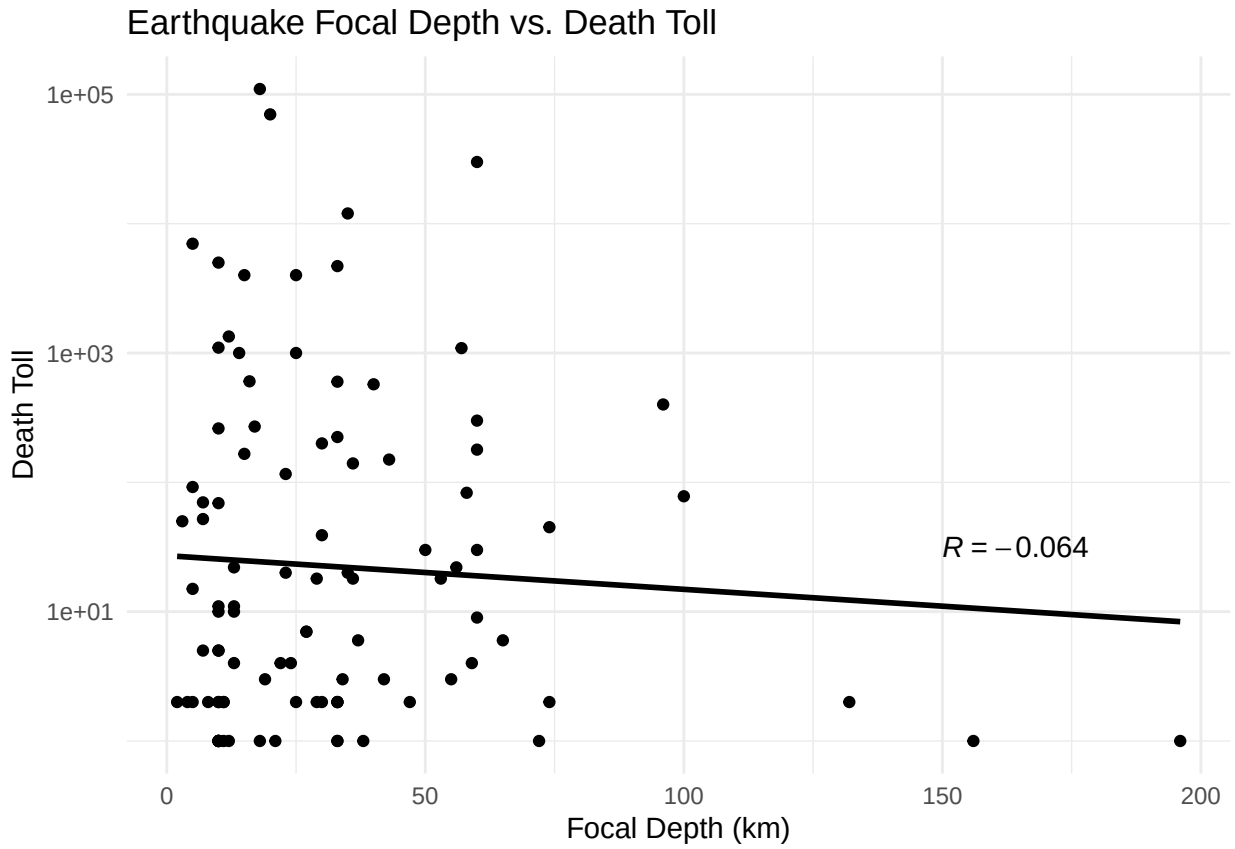
```
## i Please use ‘after_stat(r.label)‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## ‘geom_smooth()‘ using formula = ’y ~ x’

## Warning: The following aesthetics were dropped during statistical transformation: colour
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a ‘group‘ aesthetic or to convert a numerical
##   variable into a factor?
```



Scatter Plot for Focal Depth and Death Toll

```r
# Create scatter plot for focal depth vs. deaths:
ggplot(data=earthquakes_sample, aes(x=focal_depth, y=deaths)) +
  geom_point() + # scatter plot
  scale_y_log10() + # log base 10 the deaths variable due to skew
  geom_smooth(method=lm, color="black", se=FALSE) + # linear regression line with no standard error
  labs(title="Earthquake Focal Depth vs. Death Toll", x="Focal Depth (km)", y="Death Toll") + # add tit
  theme_minimal() + # add minimalist theme
  stat_cor(aes(label = ..r.label..), label.x = 150, label.y=1.5) # display Pearson correlation coeffici
```

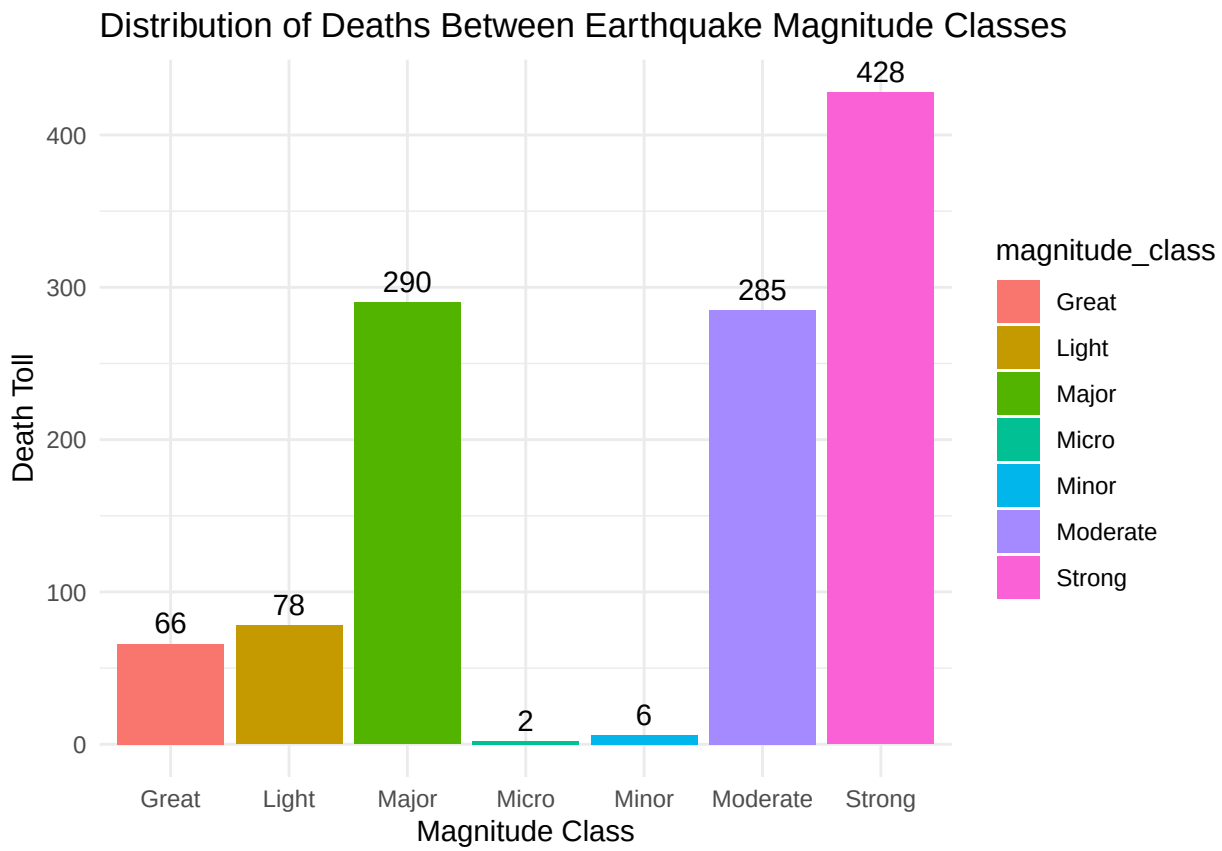
```
## `geom_smooth()` using formula = 'y ~ x'
```



Histogram for Magnitude Classes and Death Toll

```
# Magnitude class death toll:
# Create death toll variables for each magnitude class:
micro_deaths <- 0
minor_deaths <- 0
light_deaths <- 0
moderate_deaths <- 0
strong_deaths <- 0
major_deaths <- 0
great_deaths <- 0

# Iterate through dataset and add death counts to each respective magnitude
# class death toll:
for(i in 1:nrow(earthquakes)){
  if(earthquakes[i,4] == "Micro"){
    micro_deaths <- micro_deaths + 1
  }
  else if(earthquakes[i,4] == "Minor"){
    minor_deaths <- minor_deaths + 1
  }
  else if(earthquakes[i,4] == "Light"){
```

```
    light_deaths <- light_deaths + 1
  }
  else if(earthquakes[i,4] == "Moderate"){
    moderate_deaths <- moderate_deaths + 1
  }
  else if(earthquakes[i,4] == "Strong"){
    strong_deaths <- strong_deaths + 1
  }
  else if(earthquakes[i,4] == "Major"){
    major_deaths <- major_deaths + 1
  }
  else{
    great_deaths <- great_deaths + 1
  }
}

# Create vector of the resulting death tolls for each magnitude class:
death_toll_vector <- c(micro_deaths,
                       minor_deaths,
                       light_deaths,
                       moderate_deaths,
                       strong_deaths,
                       major_deaths,
                       great_deaths)

# Create a vector of the different earthquake magnitude classes:
magnitude_class_vector <- c("Micro",
                            "Minor",
                            "Light",
                            "Moderate",
                            "Strong",
                            "Major",
                            "Great")

# Create a dataframe of the earthquake magnitude classes and their respective
# death tolls:
death_toll_df <- data.frame(magnitude_class_vector, death_toll_vector)

# Rename magnitude class column in df:
death_toll_df <- death_toll_df %>% rename(magnitude_class =
                                            magnitude_class_vector)
# Rename death toll column in df:
death_toll_df <- death_toll_df %>% rename(death_toll = death_toll_vector)

# Create bar plot for the distribution of death toll based on magnitude class:
ggplot(data=death_toll_df, aes(x=magnitude_class, y=death_toll, fill=magnitude_class)) + # color the ba
  geom_bar(stat="identity") + # bar graph
  labs(title="Distribution of Deaths Between Earthquake Magnitude Classes", x="Magnitude Class", y="Dea
  theme_minimal() + # minimalist theme
  geom_text(aes(x=magnitude_class, y=death_toll, label=death_toll), vjust=-0.5) # add death toll labels
```



Discussion

I had to do some data wrangling and make some decisions that would produce more insightful visualizations. Firstly, in order to make my graphs, I took a random sample of 100 earthquakes in order to have graphs that represented a random sample of the earthquakes but weren't messy from too many data points. Then, I checked to see if there was skew in the `deaths` variable, since I suspected there would be. I found a strong right skew, which led me to the decision to use the log with base 10 of the `deaths` variable in my two scatter plots.

My first scatter plot examined the relationship between an earthquake's magnitude and its death toll. This plot showed a positive correlation between magnitude and death toll, as seen by the line of best fit and the Pearson correlation coefficient (R = 0.4) related to that line. The points in this graph are colored by magnitude class to see if there is a visible relationship between magnitude class and death toll, but not much can be determined from the coloring on this graph. This proved the need for another visualization to determine the distribition of deaths between magnitude classes. This graph will be discussed later on.

My second scatter plot visualized the relationship between an earthquake's focal depth and its death toll. This plot was a bit less direct than my first scatter plot. Firstly, we can see that most of the points are between focal depths of 0-100 km, but there are a few outliers. Within the cluster of points at focal depths of 0-100 km, the death toll varies greatly. The line of best fit conveys a weak negative correlation with a Pearson correlation coefficient of -0.064. However, this negative slope seems to be influenced more by outliers than by the overall trend in the data points.

My last visualization is a bar graph that conveys the distribution of the overall death toll between the different earthquake magnitude classes. I did not use a log scale for `deaths` in this plot because there was not such a large skew as there was in the overall dataset. In order to make this bar graph, I first had to

make a dataset of the death toll for each magnitude class. The bars in this plot are ordered by severity so it is easy to see the differences in magnitude class death toll based on earthquake severity. Based on this plot, magnitude class death toll increases in the following order:
-Micro
-Minor
-Great
-Light
-Moderate
-Major
-Strong

Overall, the three visualizations I created provide insight to answer my research question. My scatter plot of the relationship between earthquake magnitude and death toll conveys a moderate positive correlation between these two variables. My scatter plot of the relationship between earthquake focal depth conveys a weak negative correlation between earthquake focal depth and death toll, but this could be influenced by outliers and would require further analysis. Lastly, my bar plot shows that death toll does not directly increase as magnitude class severity increases and is also not uniform across magnitude classes. While the most severe magnitude class has the highest death toll, the death toll varies greatly between the rest of the magnitude classes.