Earthquake Research

Leona Pierce, Matt Collins

Intro

San Andreas is a 2015 action/thriller movie featuring Dwayne "The Rock" Johnson, who, along with his fellow costars, must survive the aftermath of the "Big One", which is a massive earthquake that has hit California (https://www.imdb.com/title/tt2126355/?ref_=ttfc_ql).

San Andreas, although not the focus of our investigation, was the motivation for the creation of our data set by FiveThirtyEight. A few days before the release of the movie, FiveThirtyEight asked 1013 Americans a series of questions relating to their knowledge and worry about the "Big One" (https://fivethirtyeight.com/features/the-rock-isnt-alone-lots-of-people-are-worried-about-the-big-one/).

Context about earthquakes

Before moving forwards, it is important to define what the "Big One" refers to. Usually, when someone refers to the "Big One", they are referencing the theoretical earthquake of a magnitude between 8-9.2 that would originate from the Cascadia subduction zone which is a 700 mile fault line off the coast of the Pacific Northwest. The likely damage of this earthquake would cover "some hundred and forty thousand square miles ... and some seven million people", or put another way, according to a director at FEMA, "everything west of Interstate 5 will be toast". This would be the second worst natural disaster in the history of North America behind the 2010 Haiti earthquake (https://www.newyorker.com/magazine/2015/07/20/the-really-big-one).

Although the "Big One" can be very concerning, earthquakes in general are actually quite common. For example, based on USGS data (https://www.kaggle.com/datasets/farazrahman/earthquake), in just the past month, there has been 9400 earthquakes in the US according to Table 1 .

Table 1: Count of earthquakes in US from 11/11 to 12/11

count	daily_count
9400	313.33

However, despite this seemingly large number of earthquakes, the earthquakes themselves actually aren't very large. Looking into this data further, the summary statistics in Table 2 show that the mean magnitude is only 1.5 on the Richter scale. Furthermore, 75% of the earthquakes are a 2 or less on the Richter scale. These numbers become even more compelling when we consider that the Richter scale is logarithmic, so a magnitude of 2 is much much less than the theoretical magnitudes of 8 or 9 that the "Big One" would have.

Table 2: Summary statistics of earthquake magnitude in US from 11/11 to 12/11

mean	median	standard_deviation	Q1	Q3	max
1.5	1.37	1.28	0.75	2	7

Set up

Despite the fact that the "Big One" would be an outlier in size, it is still very possible, so it is worthwhile to look at people's views and experiences with the "Big One". However, in order to most successfully answer our research questions below, we must first set up our environment with the appropriate packages and then import and prepare the data for analysis.

First, as far as packages, we will be using the following:

- tidyverse - This package gives us the tools that we may need for tidying data
- dplyr - This package allows us to use piping to more efficiently carry out series of functions
- janitor - This package allows us to easily create frequency tables for our data
- knitr - This package give us increased customization for our table to make them more visually appealing
- ggplot2 - This package extends R's plotting capabilities allowing us to create more detailed and visually clear plots
- usmap - This package allows use to overlay our data on a map to convey how it varies across location

The second step of preparation was to import and set up the data for analysis. Fortunately, our data was already tidied, so we did not need to do any additional work for that aspect of our data. However, one problem that we did encounter was that the original column names were the full questions. This makes the columns very unwieldy to work with, so we decided to rename the columns to names that are shorter but still convey the main idea of the given questions.

When talking about our data, it is also important to consider the FAIR and CARE principles. FINISH THIS PARAGRAPH.

Research Questions

In order to explore this data, we have selected 4 research questions to investigate.

Our first question is "How does worry differ across the United States?". We selected this question because we believe that location will affect worry since in certain areas of the US, such as the West Coast, earthquakes are more common and thus worry may be higher do to the greater knowledge of the effects of earthquakes.

Our second question is "Does knowledge about possible causes of the"Big One" affect how worried someone is?". We believe that this question could give insight into possible sources of worry, or the mitigation of it, to investigate further. If we do notice a pattern, then more research in regards to the knowledge about these possible causes would be warranted.

Our third question is "What is the relationship between a person's age, their household income, and how worried someone is about the 'Big One'?". We selected this questions to further investigate possible factors in levels of worry. Age and income are two common ways to subdivide the population along possible impact differences since factors can affect one's perspectives and thoughts about the world.

Our final question is "What is the relationship between a person's household income, if they have experienced an earthquake before, and if they have made precautions for future earthquakes?". In the first question, we looked at how region affects how much preparation one does, and we decided to do further investigation along two other possible factors in preparedness. Income affects one's capability of preparing and experience of earthquakes affects one's perception of earthquakes, so these could both be strong factors.

How does worry differ across the United States?

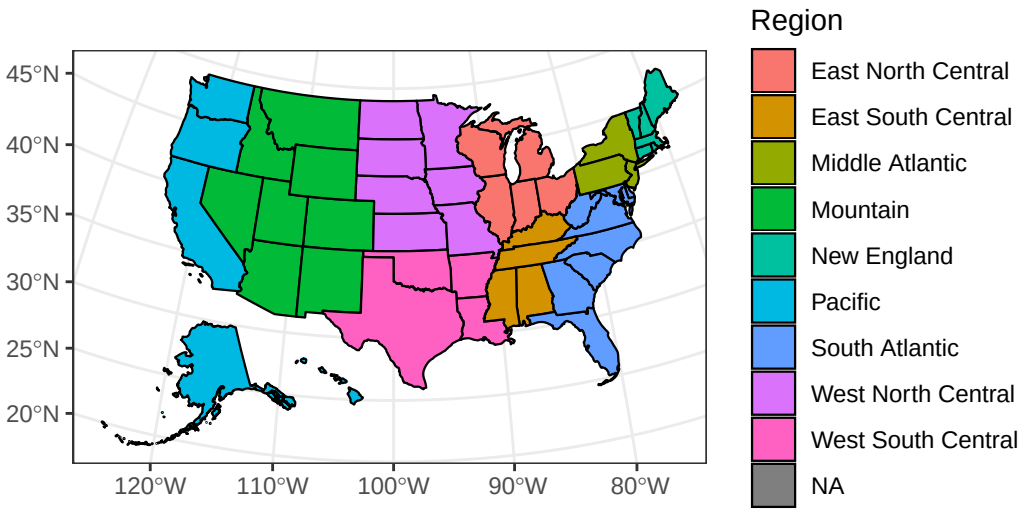
The first question we want to answer is where in the United States are people most worried about earthquakes and where in the United States are people most prepared for earthquakes? This question can be reinterpreted as the relationship between region and worriedness or preparedness.

First, let's look at where most of the responses in this data are coming from by creating a table showing the response counts of each region and the corresponding percentage of the whole.

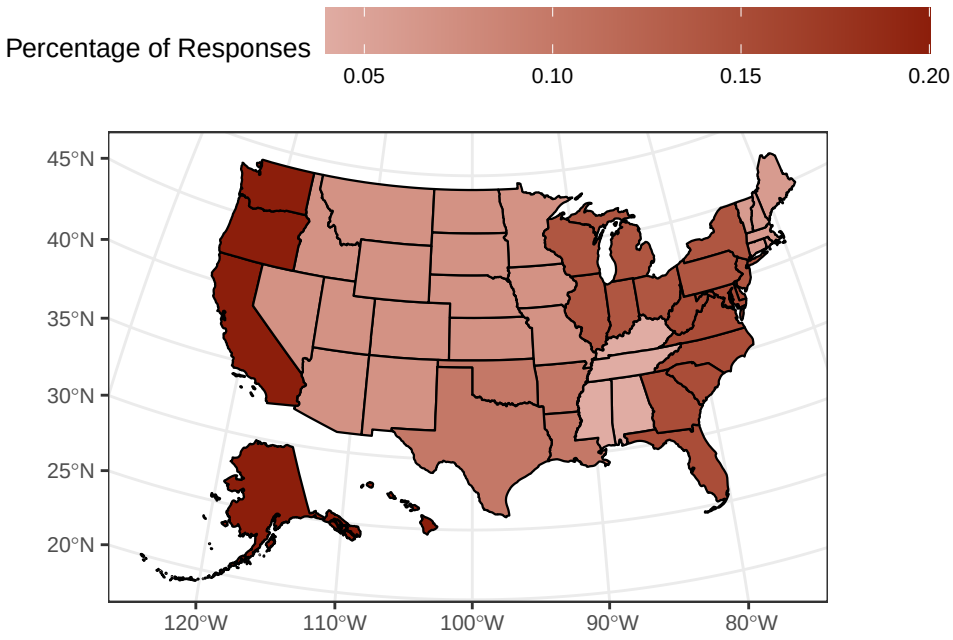
Region	Count	Percentage
	35	0.03
East North Central	140	0.14
East South Central	40	0.04
Middle Atlantic	137	0.14
Mountain	67	0.07
New England	63	0.06
Pacific	206	0.20
South Atlantic	155	0.15
West North Central	71	0.07
West South Central	99	0.10

The table shows that the responses are mostly spread out across the whole of the United States, although there are noticeably the fewest responses in the East South Central region. The blank row at the top of the table shows responses that did not give a region, which will be ignored in our analysis.

An important thing to define before we continue with our analysis is what states are contained in each region. To show maps of the US in this analysis, I will use the "usmap" package. This first map will give each region (or state in that region) a different color, and I am using the US Census Bureau's definition of the states in each region.



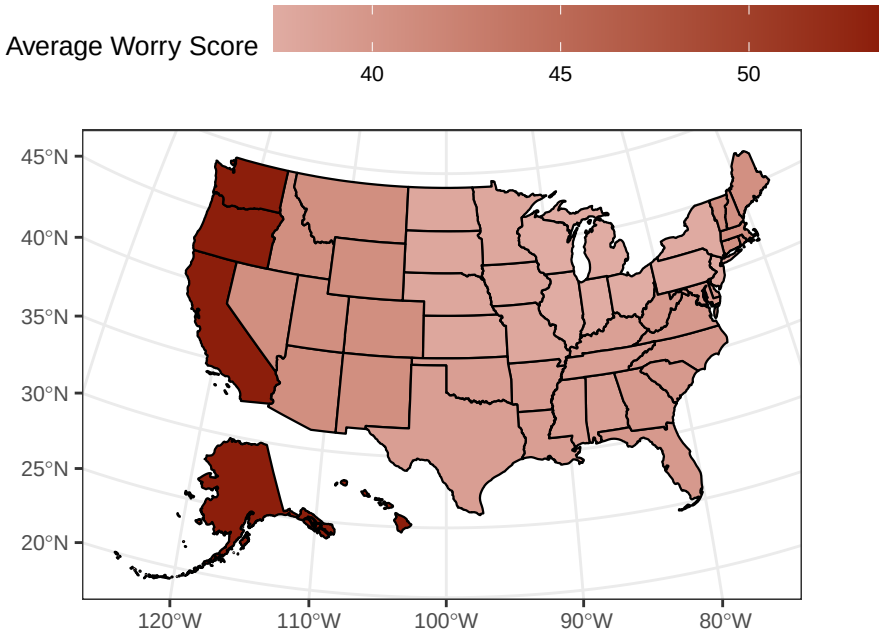
Now, let's recreate the table of regions and their respective count of responses in map form. An important thing to note about the maps in this analysis is that since we were not given an exact US state for the response, all states in a region will have identical counts or percentages.



Now I will move into answering the first part of our research question: where in the US are people most worried about earthquakes? This will be accomplished using the "general_worry" column in our dataset. First, I will create a table where each row is a combination of a US state and an average worry score. The average worry score is generated by assigning a multiplier of 1 through 5 to each response where a higher level of worry has a higher multiplier, then multiplying the region percentage of that response by its multiplier, and finally taking the average of the region total by dividing by 5 for each possible response.

```
# A tibble: 6 x 3
  Region              state avg_worry
  <chr>               <chr>     <dbl>
1 East North Central  IL         37.4
2 East North Central  IN         37.4
3 East North Central  OH         37.4
4 East North Central  MI         37.4
5 East North Central  WI         37.4
6 East South Central  KY         38.8
```

Then, we can plug this table into the same process for our response count map and get a new map showing the average level of worry, all summarized in a single map.



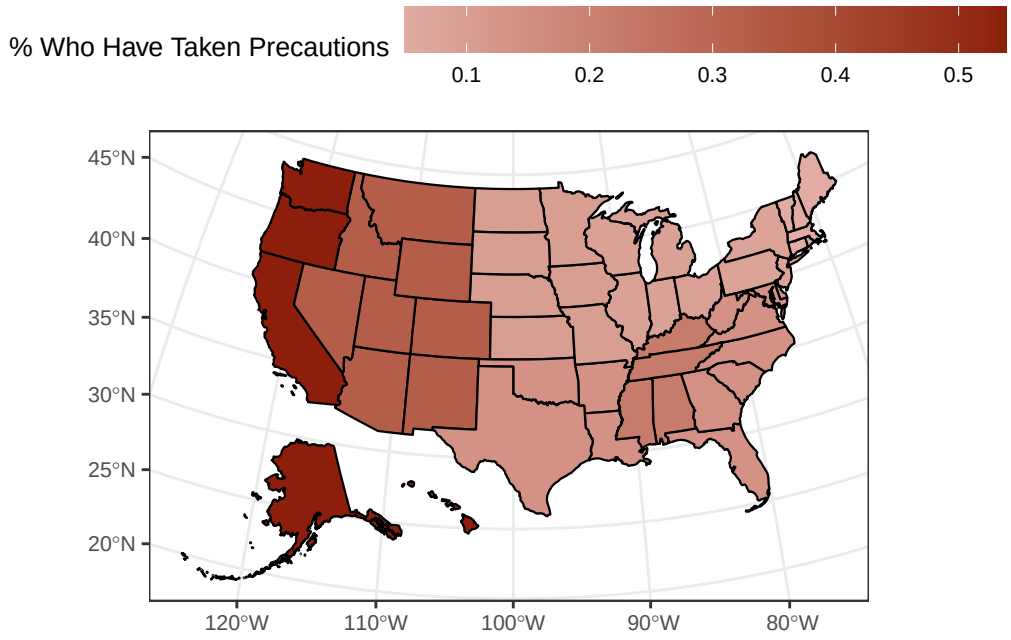
As we might have expected, the states/regions that are most worried about earthquakes are the western ones, particularly the Pacific and Mountain regions, where earthquakes are the

most common. It is important to pay attention to the scale at the top of these maps because although the shades of red make them seem very different in the level of worry, most of the regions are quite close in reality. These maps show that the western portion of the United States is generally more worried about earthquakes.

Now, we will follow this exact same process for determining where in the US people are most likely to have prepared for earthquakes, using the "taken_precautions" column. Below is the table that we will use to create the map.

```
# A tibble: 6 x 7
  Region            state taken_preacautions count total_percentage region_count
  <chr>             <chr> <chr>               <int>            <dbl>        <int>
1 East North Centr~ IL    No                    128             0.13          140
2 East North Centr~ IL    Yes                    12             0.01          140
3 East North Centr~ IN    No                    128             0.13          140
4 East North Centr~ IN    Yes                    12             0.01          140
5 East North Centr~ OH    No                    128             0.13          140
6 East North Centr~ OH    Yes                    12             0.01          140
# i 1 more variable: region_percentage <dbl>
```

Below is the map. Because it is a binary response, we only need a single map to tell us the whole story.



Just like the graphs displaying the level of worry, these graphs show that the Western part of the United States is more likely to have taken precautions for earthquakes. In fact, it almost appears to be a gradient going from west to east. Logically, this tracks because a person is more likely to have taken precautions for earthquakes if earthquakes happen relatively often. In other words, you won't prepare for something extremely unlikely to happen to you.

The following visualizations have allowed us to answer where in the US people are most worried and prepared for earthquakes, the answer being the Western part of the United States where earthquakes occur more often.

Does knowledge about possible causes of the "Big One" affect how worried someone is?

Moving beyond how worried and prepared Americans are generally about earthquakes, we wanted to look specifically at how Americans feels about the "Big One". One question we had in regard to the "Big One" is how knowledge about both the San Andreas fault line and the Yellowstone supervolcano (which are possible causes of the "Big One") may affect how worried one is. On the one hand, more knowledge may result in a greater understanding of how earthquakes occur and their effects, so there would be less irrational worry. But on the other hand, learning of the existence of these possible causes may heighten one's worry as they becoming more aware of a realistic cause of the "Big One".

The original article that this data came from, does mention this analysis in passing, but we wanted to dive deeper and actually look at the full breakdown of the data. To investigate this question, we decided to visualize the data as two way tables containing the proportions of "Yes" and "No" responses to the question "Do you think the"Big One" will occur in your lifetime?" compared to their knowledge of the fault line and super volcano in Table 4 and Table 5 respectively.

Table 4: Worry of 'Big One' occuring in one's liftime compared to knowledge of the San Andreas fault line

Knowledge level/Worry level	No	Yes	Total
	6 (0.59%)	6 (0.59%)	12 (1.18%)
Extremely familiar	55 (5.43%)	78 (7.70%)	133 (13.13%)
Not at all familiar	72 (7.11%)	34 (3.36%)	106 (10.46%)
Not so familiar	78 (7.70%)	36 (3.55%)	114 (11.25%)
Somewhat familiar	229 (22.61%)	168 (16.58%)	397 (39.19%)
Very familiar	137 (13.52%)	114 (11.25%)	251 (24.78%)
Total	577 (56.96%)	436 (43.04%)	1,013 (100.00%)

Table 5: Worry of 'Big One' occuring in one's liftime compared to knowledge of the Yellowstone supervolcano

Knowledge level/Worry level	No	Yes	Total
	6 (0.59%)	6 (0.59%)	12 (1.18%)
Extremely familiar	40 (3.95%)	53 (5.23%)	93 (9.18%)
Not at all familiar	171 (16.88%)	99 (9.77%)	270 (26.65%)
Not so familiar	142 (14.02%)	77 (7.60%)	219 (21.62%)
Somewhat familiar	152 (15.00%)	128 (12.64%)	280 (27.64%)
Very familiar	66 (6.52%)	73 (7.21%)	139 (13.72%)
Total	577 (56.96%)	436 (43.04%)	1,013 (100.00%)

Based on both tables, in most cases, more people believe that the "Big One" is not going to occur in their life time. However, the one exception is that when someone is extremely familiar with either the San Andreas fault line or the Yellowstone supervolcano, they are more likely than not to believe that the "Big One" will be in their lifetime.

This data supports the claim that more knowledge is related to more worry, but it in unclear what the underlying nature of the relationship is. We can't clearly say that more knowledge causes more worry because there are several possible other explanations. For example, the relationship could actually be the opposite such that worry causes one to seek out knowledge rather than knowledge causing one to worry.

Does age and income relate to one's worry?

In the previous section, we looked at if there are patterns between one's knowledge about fault lines and super volcanoes and their worry about the "Big One". In this section we will look at another set of variables that could have a pattern in relation to worry. To be specific, we will be looking at age and income.

In Figure 1, there are a few patterns of interest. The first is that the proportion of people who are "Not at all worried" does not appear to change very much with age and stays relatively consistent. This suggests that age is not a factor in one having very low worry levels about the "Big One".

The second pattern is that the proportion of people that are "Extremely worried" decreases as age increase while the proportion of people that are "Not so worried" increases as age increases. This in some ways makes sense because in order for proportionally less people to have one level of worry, there needs to be proportionally more people to have other levels of worry to account for this change.

Another interesting thing of note in regards to the second pattern is that the "Very worried" and "Somewhat worried" are approximately the same proportion across all age groups. This

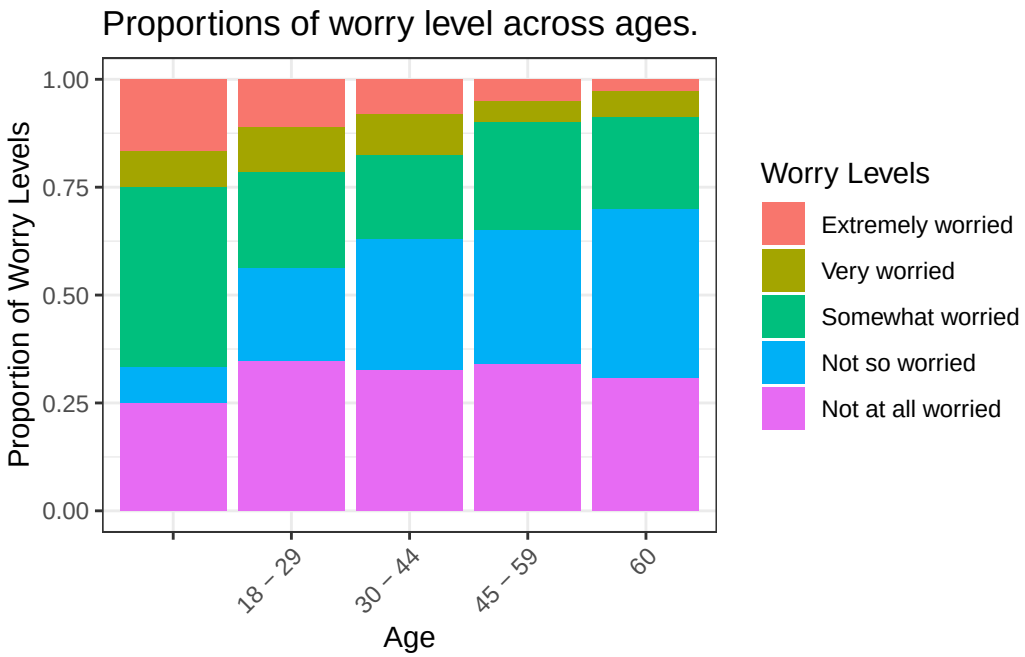


Figure 1: Proportions of worry level across ages.

is likely the result of one of two changes. The first is that the decrease in "Extremely" and increase in "Not so" is a direct shift from the first to the second. The second explanation is that there us a shift where many people decrease their worry across "Extremely", "Very", "Somewhat" and "Not so", but the proportions of these changes were equivalent such that as a net change it merely appears as a decrease in "Extremely" and an increase in "Not so". In order to more accurately determine this, we would need to conduct this survey multiple times across the same group of people to see how individual answers change with age.

As compared to Figure 1, the patterns in Figure 2 are less apparent. One initial pattern that can be identified is that there seems to be an overall decrease in worry as income increases, but the $200,000 and up income range does not follow this trend. However, this could be the result of outliers due to the fact that this income bracket is uncapped.

Beyond this initial trend, any other conclusions would likely require greater information and more data gathering; however, another possible route for further investigation would be an analysis of how these same demographic attributes affect another element of worriedness about earthquakes. Specifically, we will look at the effects on preparations and precautions in regards to earthquakes.

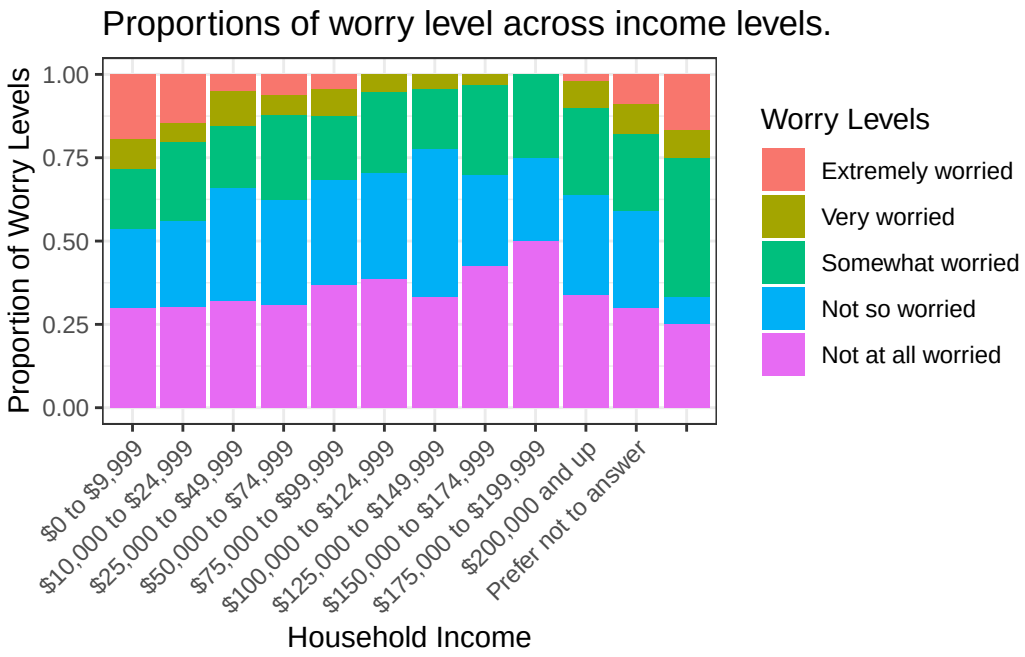


Figure 2: Proportions of worry level across income levels.

Who is prepping for earthquakes?

Conclusion

```
library(tidyverse)
library(janitor)
library(dplyr)
library(knitr)
library(ggplot2)
library(usmap)
earthquake_monthly <- read.csv("./data/earthquake_monthly.csv")

monthly_count <- earthquake_monthly %>%
  summarise(
    count = n(),
    daily_count = n()/30
  )

monthly_count %>%
  kable(
```

```
    digits = 2  # Round to 2 decimal places
  )
monthly_summary <- earthquake_monthly %>%
  summarise(
    mean = mean(mag, na.rm = TRUE),
    median = median(mag, na.rm = TRUE),
    standard_deviation = sd(mag, na.rm = TRUE),
    Q1 = quantile(mag, 0.25, na.rm = TRUE),
    Q3 = quantile(mag, 0.75, na.rm = TRUE),
    max = max(mag, na.rm = TRUE)
  )

monthly_summary %>%
  kable(
    digits = 2  # Round to 2 decimal places
  )
earthquake <- read.csv("./data/earthquake_data.csv")

names(earthquake)[1] <- "general_worry"
names(earthquake)[2] <- "big_one_worry"
names(earthquake)[3] <- "big_one_occur"
names(earthquake)[4] <- "experienced_earthquake"
names(earthquake)[5] <- "taken_preacautions"
names(earthquake)[6] <- "san_andreas_familiar"
names(earthquake)[7] <- "yellowstone_familiar"
names(earthquake)[8] <- "age"
names(earthquake)[9] <- "gender"
names(earthquake)[10] <- "household_income"
names(earthquake)[11] <- "region"

earthquake <- earthquake

region_counts <- earthquake %>% count(region) %>%
  mutate(Percentage = round(n / sum(n), 2)) %>%
  rename(Region = region, Count = n)
region_counts %>% kable
state_to_region <- list(
  "East North Central" = c("IL", "IN", "OH", "MI", "WI"),
  "East South Central" = c("KY", "TN", "MS", "AL"),
  "Middle Atlantic" = c("NY", "PA", "NJ"),
  "Mountain" = c("MT", "ID", "WY", "CO", "NM", "UT", "AZ", "NV"),
  "New England" = c("ME", "VT", "NH", "MA", "CT", "RI"),
```

```
  "Pacific" = c("CA", "OR", "WA", "HI", "AK"),
  "South Atlantic" = c("WV", "MD", "DE", "VA", "NC", "SC", "GA", "FL"),
  "West North Central" = c("ND", "SD", "MN", "IA", "MO", "NE", "KS"),
  "West South Central" = c("AR", "TX", "OK", "LA")
)

state_region_df <- enframe(state_to_region, name = "Region", value = "state") %>% unnest(sta

state_counts <- state_region_df %>%
  left_join(region_counts, by = "Region")
plot_usmap(data = state_counts,
           values = "Region",
           theme = theme_bw())
plot_usmap(data = state_counts,
           values = "Percentage",
           theme = theme_bw()) +
  scale_fill_gradient(
    labels = scales::label_number(big.mark = ','),
    breaks = c(0.05, 0.10, 0.15, 0.20),
    high = '#8c1e0b',
    low = '#e0aca3'
  ) +
  theme(
    text = element_text(size = 10),
    legend.position = 'top'
  ) +
  labs(fill = 'Percentage of Responses') +
  guides(
    fill = guide_colorbar(
      barwidth = unit(8, 'cm')
    )
  )
region_worry <- earthquake %>% count(region, general_worry) %>%
  mutate(total_percentage = round(n / sum(n), 2)) %>%
  group_by(region) %>%
  mutate(region_count = sum(n),
         region_percentage = round(n / region_count, 2)) %>%
  rename(Region = region, count = n) %>%
  mutate(region_score = ifelse(general_worry == "Extremely worried", region_percentage * 500

region_avg_worry <- region_worry %>% group_by(Region) %>%
                    summarize(avg_worry = sum(region_score) / 5)
```

```r
state_worry <- state_region_df %>%
  left_join(region_avg_worry, by = "Region")

head(state_worry)
plot_usmap(data = state_worry,
           values = "avg_worry",
           theme = theme_bw()) +
  scale_fill_gradient(
    high = '#8c1e0b',
    low = '#e0aca3'
  ) +
  theme(
    text = element_text(size = 10),
    legend.position = 'top'
  ) +
  labs(fill = 'Average Worry Score') +
  guides(
    fill = guide_colorbar(
      barwidth = unit(8, 'cm')
    )
  )
region_precautions <- earthquake %>% count(region, taken_preacautions) %>%
  mutate(total_percentage = round(n / sum(n), 2)) %>%
  group_by(region) %>%
  mutate(region_count = sum(n),
         region_percentage = round(n / region_count, 2)) %>%
  rename(Region = region, count = n)

state_precautions <- state_region_df %>%
  left_join(region_precautions, by = "Region")

head(state_precautions)
plot_usmap(data = state_precautions %>% filter(taken_preacautions == "Yes"),
           values = "region_percentage",
           theme = theme_bw()) +
  scale_fill_gradient(
    high = '#8c1e0b',
    low = '#e0aca3'
  ) +
  theme(
    text = element_text(size = 10),
    legend.position = 'top'
```

```
  ) +
  labs(fill = '% Who Have Taken Precautions') +
  guides(
    fill = guide_colorbar(
      barwidth = unit(8, 'cm')
    )
  )

san_andreas_table <- earthquake %>%
  tabyl(san_andreas_familiar, big_one_occur) %>%
  adorn_totals(where = c("row", "col") ) %>%
  adorn_percentages(denominator = "all") %>%
  adorn_pct_formatting(digits = 2) %>%
  adorn_title(
  placement = "combined",
  row_name = "Knowledge level",
  col_name = "Worry level")

san_andreas_formatNs <- attr(san_andreas_table, "core") %>%
adorn_totals(where = c("row", "col")) %>%
mutate(
across(where(is.numeric), format, big.mark = ",")
)
san_andreas_FreqTab <- san_andreas_table %>%
adorn_ns(position = "front", ns = san_andreas_formatNs)

san_andreas_FreqTab %>% kable(digits = c(0, 0, 2, 2, 2))

yellowstone_table <- earthquake %>%
  tabyl(yellowstone_familiar, big_one_occur) %>%
  adorn_totals(where = c("row", "col") ) %>%
  adorn_percentages(denominator = "all") %>%
  adorn_pct_formatting(digits = 2) %>%
  adorn_title(
  placement = "combined",
  row_name = "Knowledge level",
  col_name = "Worry level")

yellowstone_formatNs <- attr(yellowstone_table, "core") %>%
adorn_totals(where = c("row", "col")) %>%
mutate(
across(where(is.numeric), format, big.mark = ",")
```

```
)
yellowstone_FreqTab <- yellowstone_table %>%
adorn_ns(position = "front", ns = yellowstone_formatNs)

yellowstone_FreqTab %>% kable(digits = c(0, 0, 2, 2, 2))

age_worry_proportion <- earthquake %>%
  group_by(age, big_one_worry) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count))

worry_order <- c("Extremely worried",
                 "Very worried",
                 "Somewhat worried",
                 "Not so worried",
                 "Not at all worried")

age_worry_proportion$big_one_worry <- factor(
      age_worry_proportion$big_one_worry,
      levels = worry_order)

ggplot(age_worry_proportion, aes(x = age,
                                 y = proportion,
                                 fill = big_one_worry)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Age",
       y = "Proportion of Worry Levels",
       fill = "Worry Levels",
       title = "Proportions of worry level across ages.") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
income_worry_proportion <- earthquake %>%
  group_by(household_income, big_one_worry) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count))

worry_order <- c("Extremely worried",
                 "Very worried",
                 "Somewhat worried",
                 "Not so worried",
                 "Not at all worried")
```

```r
income_worry_proportion$big_one_worry <- factor(
  income_worry_proportion$big_one_worry,
  levels = worry_order)

x_order <- c("$0 to $9,999",
             "$10,000 to $24,999",
             "$25,000 to $49,999",
             "$50,000 to $74,999",
             "$75,000 to $99,999",
             "$100,000 to $124,999",
             "$125,000 to $149,999",
             "$150,000 to $174,999",
             "$175,000 to $199,999",
             "$200,000 and up",
             "Prefer not to answer",
             "")

income_worry_proportion$household_income <- factor(
  income_worry_proportion$household_income,
  levels = x_order)

ggplot(income_worry_proportion, aes(x = household_income,
                                    y = proportion,
                                    fill = big_one_worry)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(x = "Household Income",
       y = "Proportion of Worry Levels",
       fill = "Worry Levels",
       title = "Proportions of worry level across income levels.") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```