Project title

Exploratory data analysis

marvelous echidna

Research question(s)

Are there areas that are more prone to experiencing landslides? If so, what types of areas are more vulnerable to landslides? Are landslides more deadly in different areas around the world? Have landslides increased in frequency?

Data collection and cleaning

The data was collected by searching for news reports, scientific reports, eyewitness statements, aerial photography, as well as other media that reliably reported the details of the landslide event (source: https://doi.org/10.1007/s11069-009-9401-4). The database records each landslide event that has been reported since 2007 until March 2016. The database records observations such as the date, time, trigger, fatalities, geographic location, as well as a host of other useful information. As of right now, we have changed the columns event_date into a month date year attribute and made new columns date, time and AM/PM using :

"data <- data |> separate_wider_delim(cols = event_date, delim = ' ', names = c('date','time','AM/PM')) |> mutate(date = mdy(date))"

Then we removed columns that aren't really necessary right now:

"data <- data |> select(-source_link, -photo_link, -notes, -event_import_source, -event_import_id, -storm_name, -event_title)"

Data description

The dataset used for this analysis contains information on landslide events from various countries around the world. The dataset comprises 11,033 observations and includes the following key variables:

source_name: to identify news articles event_id: A unique identifier for each landslide event. date: The date on which the landslide event occurred, in the format "YYYY-MM-DD". time and AM/PM: The time of day when the landslide event took place, with separate columns for the time (in the format "HH:MM:SS") and the AM/PM indicator. event_description: A brief description of the landslide event, typically including the location and any notable details. location_description: A more detailed description of the location where the landslide occurred, often including the village, county, province, and other geographic identifiers. location_accuracy: An indication of the accuracy or precision of the location information provided. landslide_category: The category or type of landslide event, such as "landslide," "mudslide," or others. landslide_trigger: The trigger or cause of the landslide event, such as "rain," "downpour," "monsoon," or others. landslide_size: The size or scale of the landslide event, categorized as "large," "small," "medium," or others. fatality_count: The number of fatalities or deaths resulting from the landslide event. country_name and country_code: The name and code of the country where the landslide event occurred. admin_division_name and admin_division_population: The name and population of the administrative division (e.g., state, province) affected by the landslide event. gazeteer_closest_point and gazeteer_distance: The closest geographic point and its distance from the landslide location, based on a gazetteer or geographic reference. submitted_date and created_date: The dates when the landslide event was submitted and created in the database, respectively. last_edited_date: The date when the landslide event entry was last edited or updated. longitude and latitude: The geographic coordinates (longitude and latitude) of the landslide event location. The dataset covers landslide events from 1988 to now, allowing for the analysis of spatial and temporal patterns, as well as the identification of potential factors contributing to landslide occurrences and their associated impacts.

Data limitations

The dataset appears to have landslide events from various countries, but it's unclear if the coverage is comprehensive or if there are geographic biases. Some regions of the world may be underrepresented or overrepresented, which could skew the analysis of areas prone to landslides and their associated fatalities.

The dataset seems to cover a specific time range (based on the date column), but the exact range is not apparent from the glimpse. If the time period is relatively short or does not cover a sufficiently long timeframe, it may be difficult to reliably analyze trends and changes in landslide frequency over time.

Several columns, such as injury_count, landslide_setting, and admin_division_population, have many missing values (indicated by NA). This could limit the analysis and lead to potential biases if the missing data is not random or if it is concentrated in specific regions or landslide types.

To accurately compare landslide fatalities across different areas and populations, it may be necessary to normalize the fatality counts by population size or other relevant factors. The dataset doesn't seem to include population data for all locations, which could make such normalization challenging.

While the dataset includes columns like landslide_trigger, landslide_size, and landslide_category, it's unclear how these characteristics are defined and measured consistently across different events and sources. Inconsistencies in classification could lead to inaccurate comparisons.

Exploratory data analysis

Perform an (initial) exploratory data analysis.

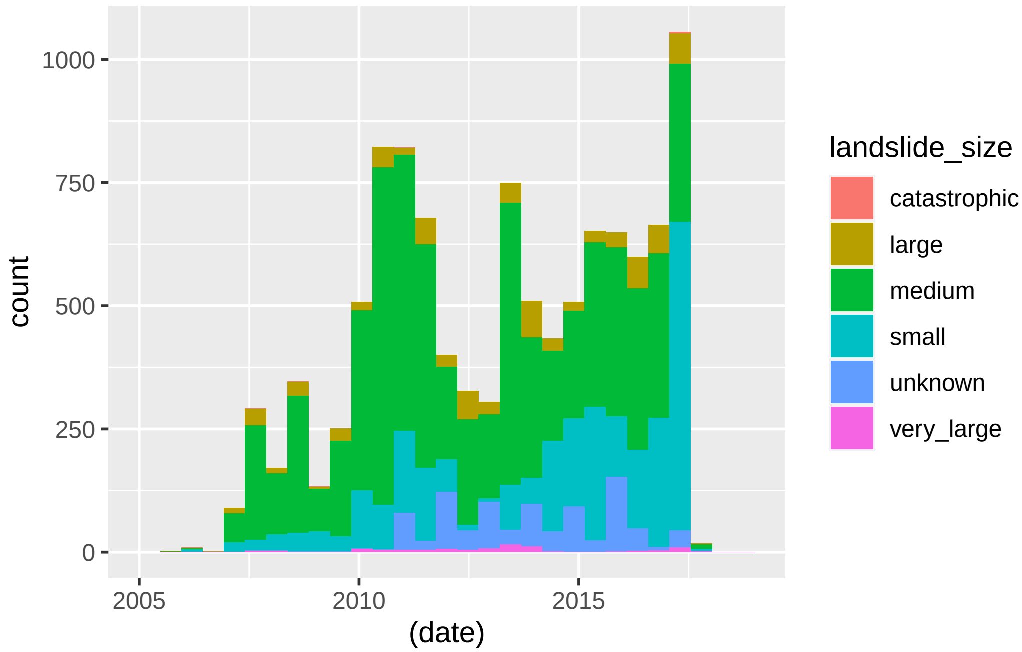
```r
library(tidyverse)

data <- read_csv('data/Global_Landslide_Catalog_Export.csv')
colnames(data)
```

```
 [1] "source_name"               "source_link"
 [3] "event_id"                  "event_date"
 [5] "event_time"                "event_title"
 [7] "event_description"         "location_description"
 [9] "location_accuracy"         "landslide_category"
[11] "landslide_trigger"         "landslide_size"
[13] "landslide_setting"         "fatality_count"
[15] "injury_count"              "storm_name"
[17] "photo_link"                "notes"
[19] "event_import_source"       "event_import_id"
[21] "country_name"              "country_code"
[23] "admin_division_name"       "admin_division_population"
[25] "gazeteer_closest_point"    "gazeteer_distance"
[27] "submitted_date"            "created_date"
[29] "last_edited_date"          "longitude"
[31] "latitude"
```

```r
data <- data |>
  separate_wider_delim(cols = event_date, delim = ' ',
                       names = c('date','time','AM/PM')) |>
  mutate(date = mdy(date))
```

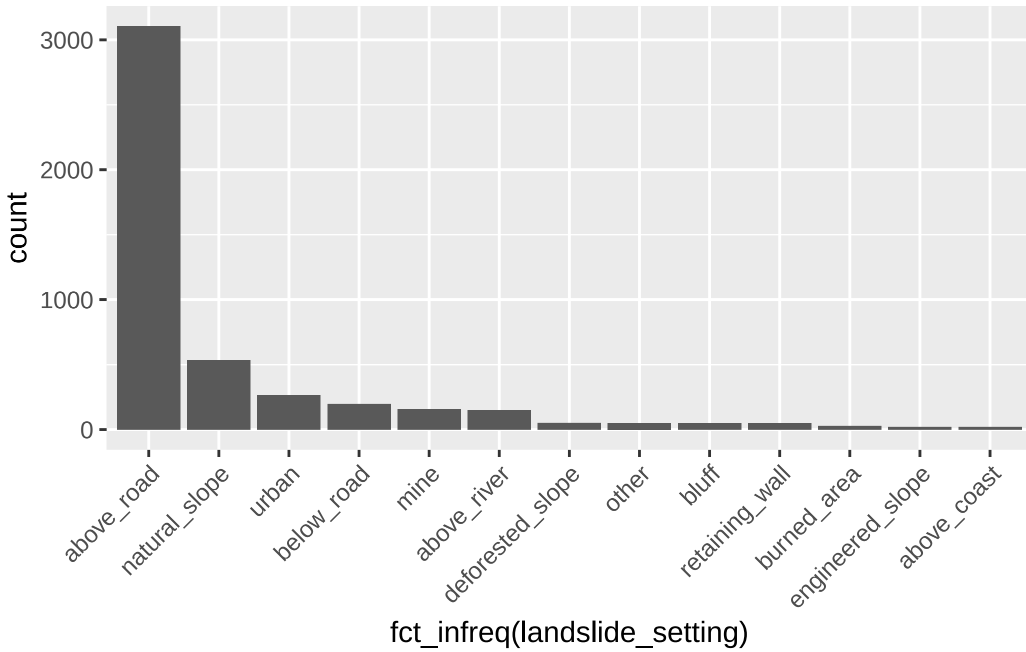
```r
data <- data |>
  select(-source_link, -photo_link, -notes, -event_import_source,
         -event_import_id, -storm_name, -event_title)

data |>
  drop_na(landslide_size) |>
  ggplot(aes(x = (date), fill = landslide_size)) +
  geom_histogram() +
  xlim(c(make_date(year = 2005),make_date(year = 2019)))
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

Warning: Removed 30 rows containing non-finite values (`stat_bin()`).

Warning: Removed 12 rows containing missing values (`geom_bar()`).
```



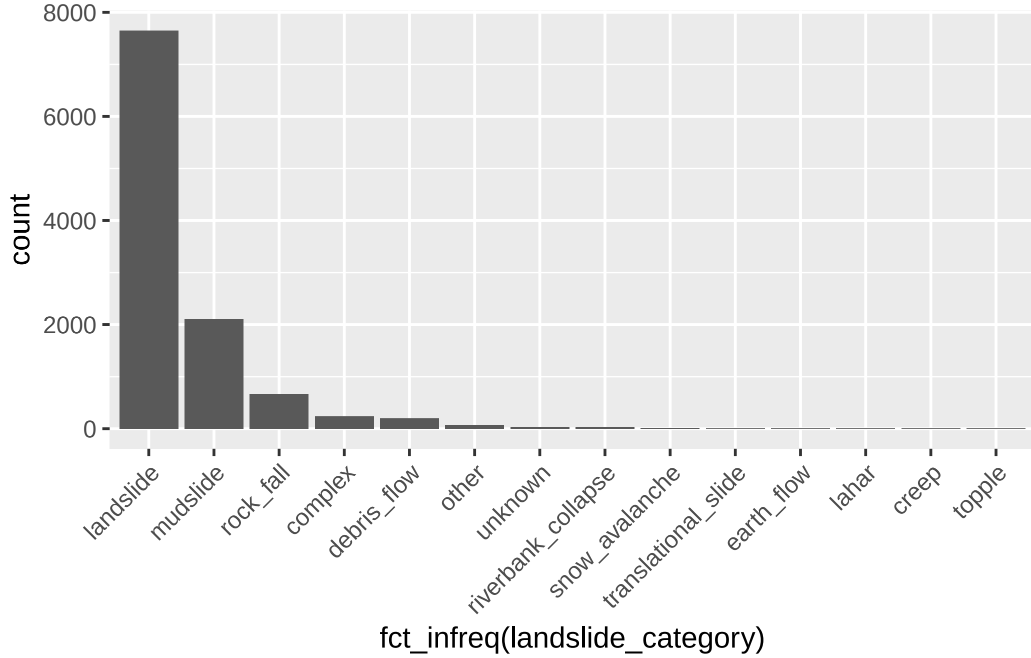

```r
data |>
  drop_na(landslide_setting) |>
  filter(landslide_setting != 'unknown') |>
  ggplot(aes(x = fct_infreq(landslide_setting))) +
  geom_histogram(stat = 'count') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
`binwidth`, `bins`, and `pad`
```



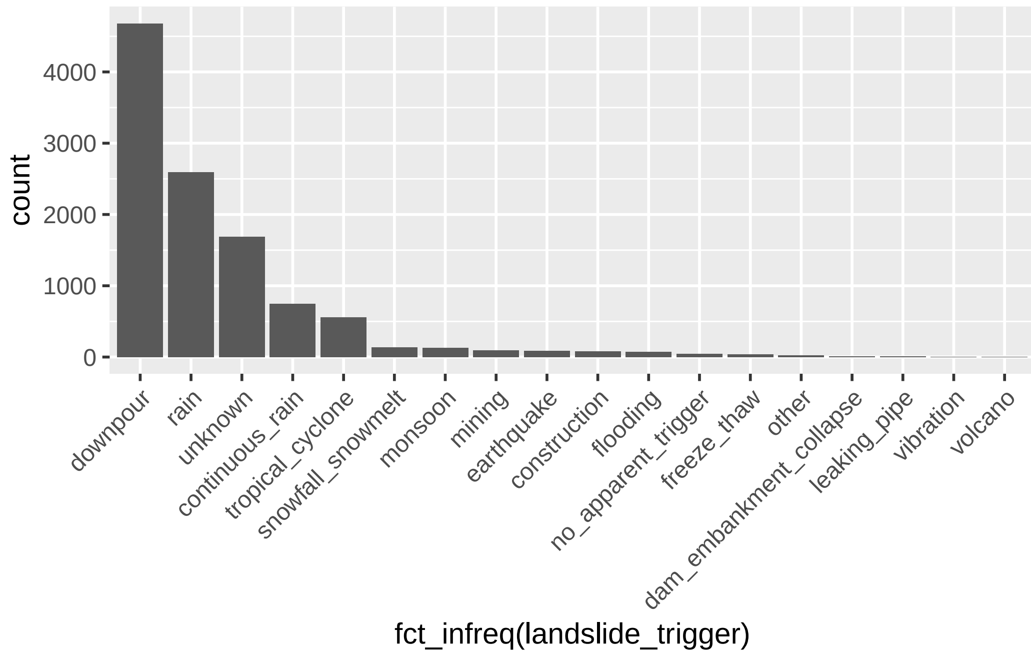

```r
data |>
    drop_na(landslide_category) |>
    ggplot(aes(x = fct_infreq(landslide_category))) +
    geom_histogram(stat = 'count') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
`binwidth`, `bins`, and `pad`
```



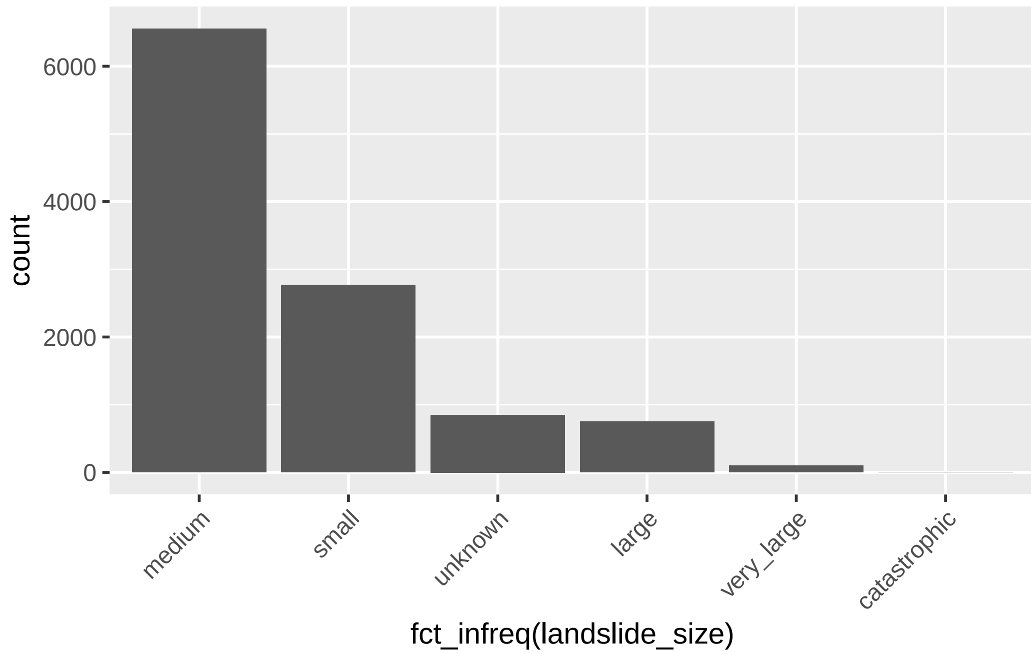

```r
data |>
    drop_na(landslide_trigger) |>
    ggplot(aes(x = fct_infreq(landslide_trigger))) +
    geom_histogram(stat = 'count') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
`binwidth`, `bins`, and `pad`
```



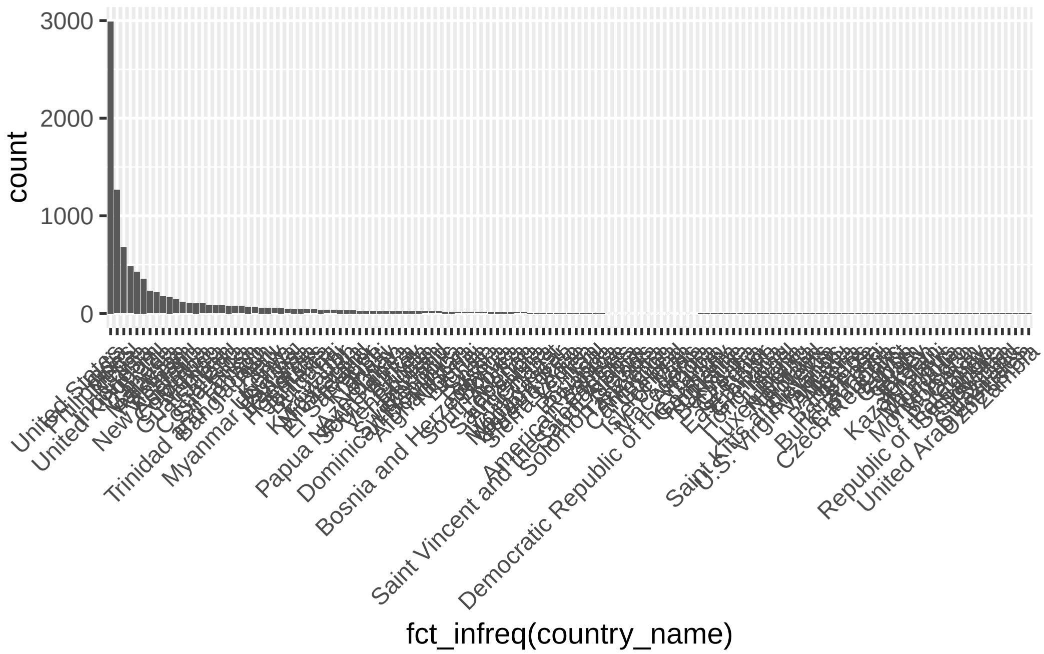

```r
data |>
    drop_na(landslide_size) |>
    ggplot(aes(x = fct_infreq(landslide_size))) +
    geom_histogram(stat = 'count') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
`binwidth`, `bins`, and `pad`
```

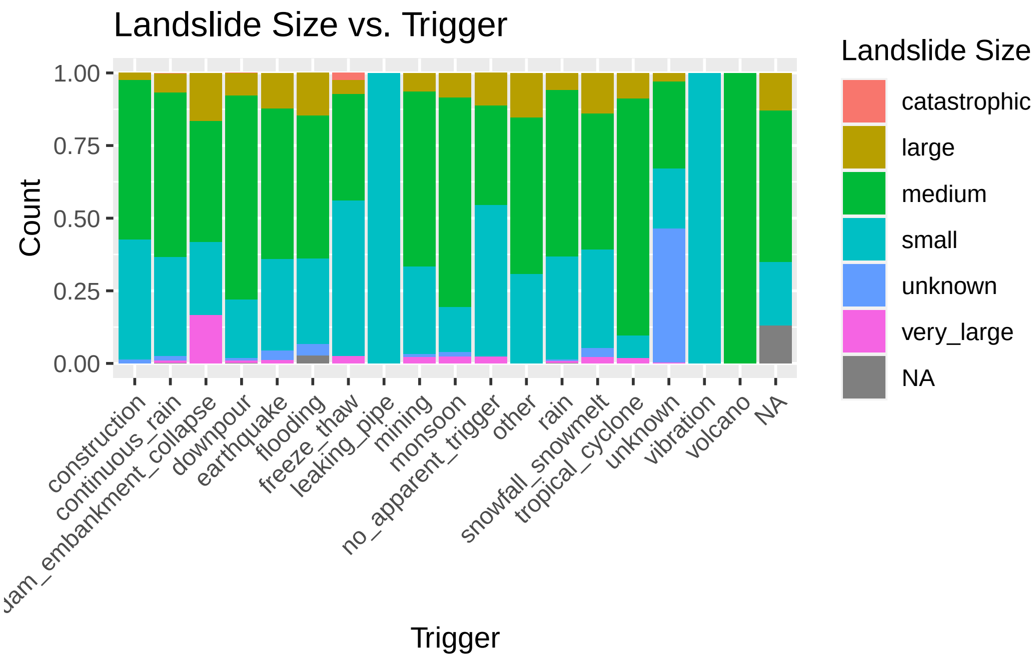



```
data |>
    drop_na(country_name) |>
    ggplot(aes(x = fct_infreq(country_name))) +
    geom_histogram(stat = 'count') +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

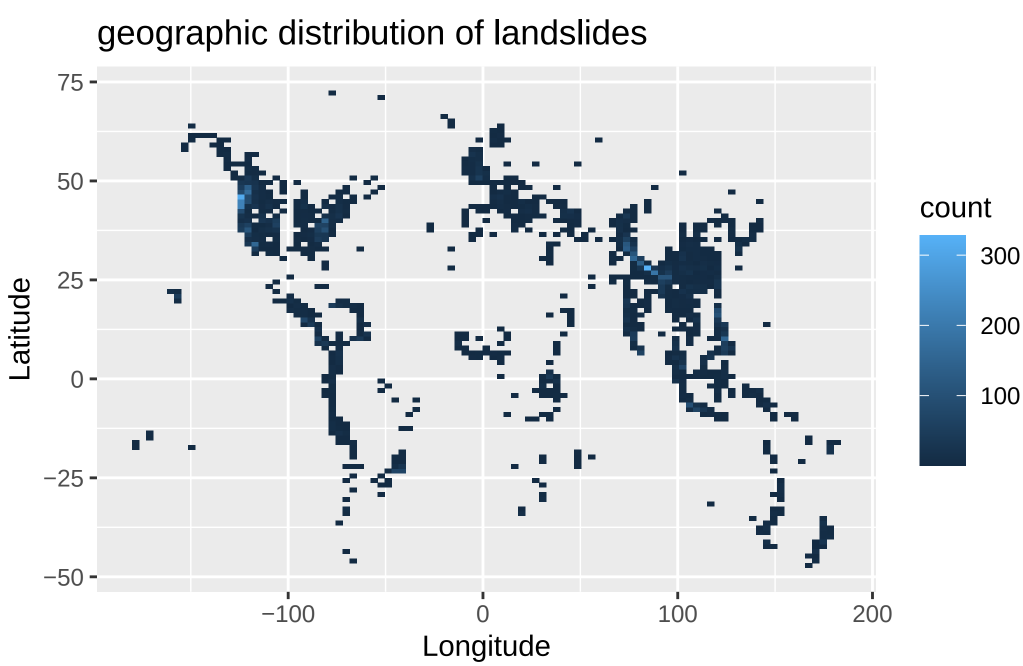
```
Warning in geom_histogram(stat = "count"): Ignoring unknown parameters:
`binwidth`, `bins`, and `pad`
```



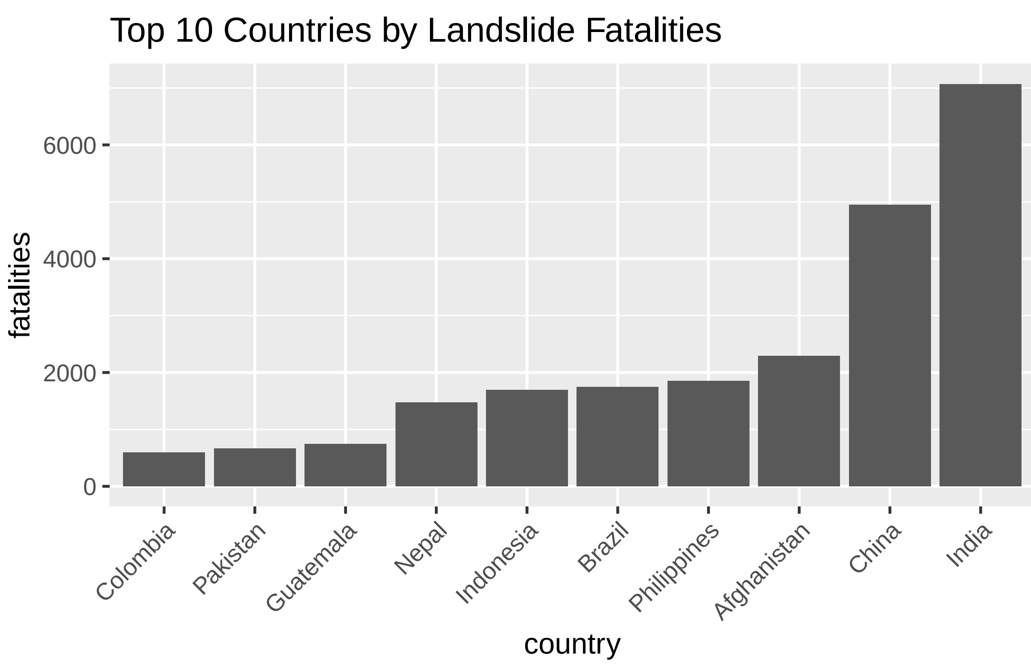

```r
ggplot(data, aes(x = landslide_trigger, fill = landslide_size)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Landslide Size vs. Trigger",
       x = "Trigger",
       y = "Count",
       fill = "Landslide Size")
```




```r
ggplot(data, aes(x = longitude, y = latitude)) +
  geom_bin2d(bins = 100) +
  labs(title = "geographic distribution of landslides",
       x = "Longitude", y = "Latitude")
```

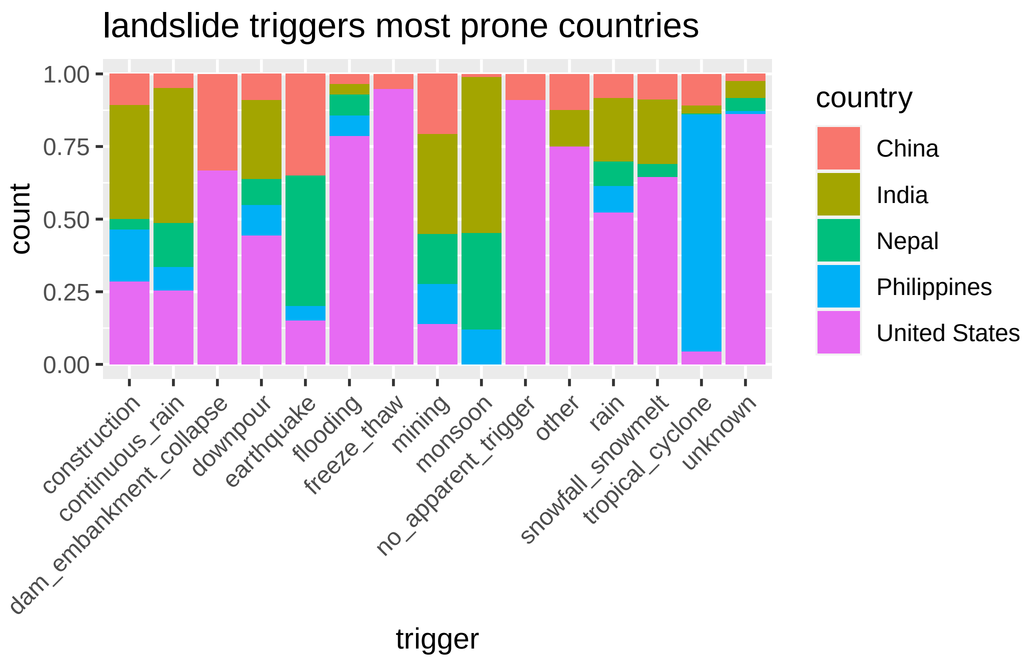


```r
data |>
  filter(!is.na(country_name)) |>
  group_by(country_name) |>
  summarise(total_fatalities = sum(fatality_count, na.rm = TRUE)) |>
  top_n(10, total_fatalities) |>
  ggplot(aes(
    x = reorder(country_name, total_fatalities),
    y = total_fatalities)) +
  geom_col() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))+
  labs(
    title = "Top 10 Countries by Landslide Fatalities",
    x = "country",
    y = "fatalities")
```




```r
top_countries <- data |>
  filter(!is.na(country_name)) |>
  count(country_name) |>
  top_n(5, n) |>
  pull(country_name)

data |>
  filter(country_name %in% top_countries) |>
```

```r
  ggplot(aes(x = landslide_trigger, fill = country_name)) +
  geom_bar(position = "fill") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "landslide triggers most prone countries",
       x = "trigger",
       y = "count",
       fill = "country")
```




```r
# landslide count by category
total_by_category <- data |>
  count(landslide_category)
total_by_category
```

```
# A tibble: 15 x 2
   landslide_category      n
   <chr>               <int>
 1 complex               232
 2 creep                   5
 3 debris_flow           194
 4 earth_flow              7
 5 lahar                   7
 6 landslide            7648
```

```
 7 mudslide              2100
 8 other                   68
 9 riverbank_collapse      37
10 rock_fall              671
11 snow_avalanche          15
12 topple                   1
13 translational_slide      9
14 unknown                 38
15 <NA>                     1
```

```r
#avg deaths by country
avg_fatalities_by_country <- data |>
  group_by(country_name) |>
  summarise(average_fatalities = mean(fatality_count, na.rm = TRUE))
avg_fatalities_by_country
```

```
# A tibble: 142 x 2
   country_name   average_fatalities
   <chr>                       <dbl>
 1 Afghanistan              191.
 2 Albania                    0
 3 Algeria                    6
 4 American Samoa             0
 5 Angola                     0
 6 Argentina                  1.5
 7 Armenia                    0.75
 8 Australia                  0.0222
 9 Austria                    0.5
10 Azerbaijan                 0.2
# i 132 more rows
```

```r
# Distribution of landslides by geographic setting
distribution_by_setting <- data |>
  count(landslide_setting)
distribution_by_setting
```

```
# A tibble: 15 x 2
   landslide_setting     n
   <chr>             <int>
 1 above_coast          20
 2 above_river         149
```

```
 3 above_road         3104
 4 below_road          199
 5 bluff                48
 6 burned_area          28
 7 deforested_slope     53
 8 engineered_slope     22
 9 mine                157
10 natural_slope       531
11 other                50
12 retaining_wall       48
13 unknown            6291
14 urban               264
15 <NA>                 69
```

```r
# Population in admin divisions affected by landslides
affected_population <- data |>
  group_by(landslide_setting) |>
  summarise(average_population = mean(admin_division_population, na.rm = TRUE))
affected_population
```

```
# A tibble: 15 x 2
   landslide_setting average_population
   <chr>                          <dbl>
 1 above_coast                     NaN
 2 above_river                   24802.
 3 above_road                    34341.
 4 below_road                    42865.
 5 bluff                         27063.
 6 burned_area                   32071.
 7 deforested_slope              47429.
 8 engineered_slope             120083
 9 mine                         190643.
10 natural_slope                 82960.
11 other                         70593.
12 retaining_wall               476343.
13 unknown                      188815.
14 urban                        929712.
15 <NA>                            NaN
```

```r
#median # of deaths by landslide size
median_fatalities_by_size <- data |>
  group_by(landslide_size) |>
```

```r
  summarise(median_fatalities = median(fatality_count, na.rm = TRUE))

median_fatalities_by_size
```

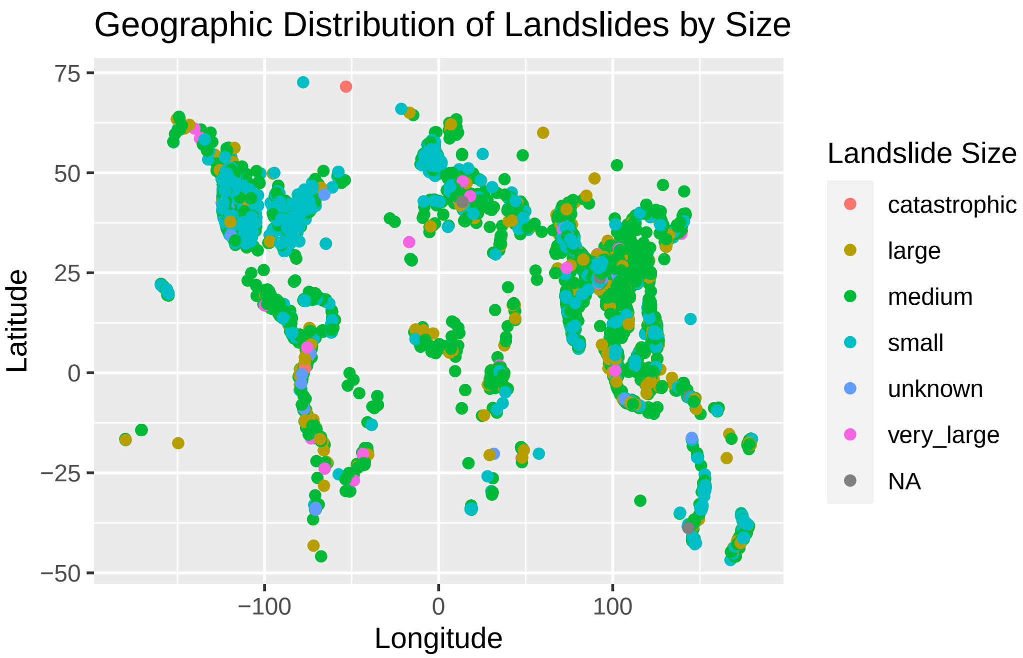
```
# A tibble: 7 x 2
  landslide_size median_fatalities
  <chr>                      <dbl>
1 catastrophic                 103
2 large                          3
3 medium                         0
4 small                          0
5 unknown                        0
6 very_large                     8
7 <NA>                           0
```

```r
# Proportion of Landslides by Trigger Type
landslide_trigger_proportion <- data |>
  count(landslide_trigger) |>
  mutate(proportion = n / sum(n)) |>
  select(landslide_trigger, proportion)

landslide_trigger_proportion
```

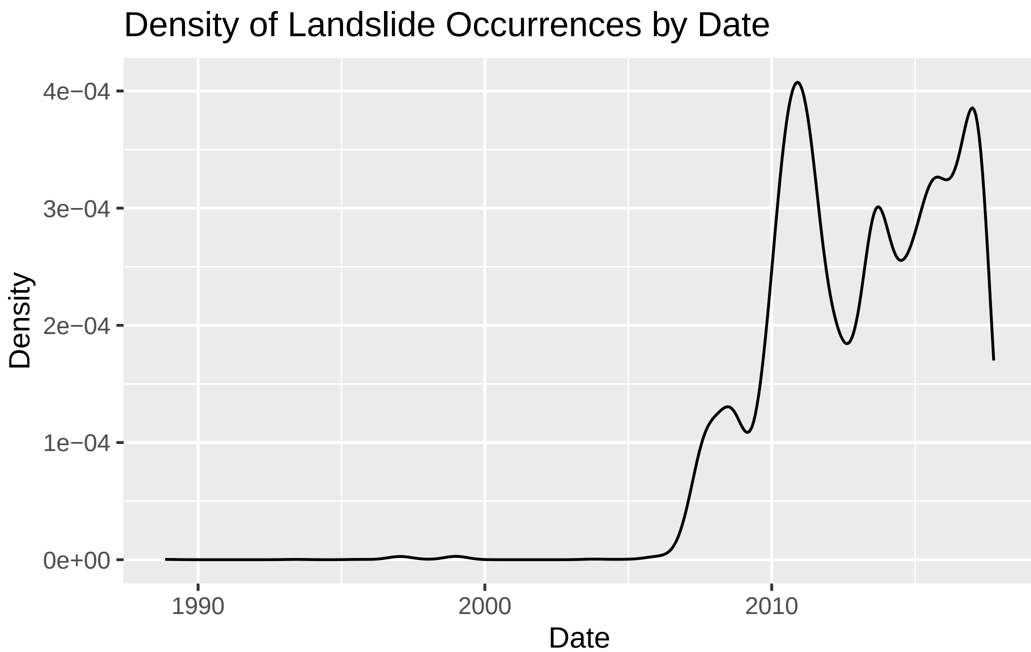
```
# A tibble: 19 x 2
   landslide_trigger       proportion
   <chr>                        <dbl>
 1 construction              0.00743
 2 continuous_rain           0.0678
 3 dam_embankment_collapse   0.00109
 4 downpour                  0.424
 5 earthquake                0.00807
 6 flooding                  0.00680
 7 freeze_thaw               0.00372
 8 leaking_pipe              0.000906
 9 mining                    0.00843
10 monsoon                   0.0117
11 no_apparent_trigger       0.00399
12 other                     0.00236
13 rain                      0.235
14 snowfall_snowmelt         0.0122
15 tropical_cyclone          0.0508
```

```
16 unknown                     0.153
17 vibration                   0.0000906
18 volcano                     0.0000906
19 <NA>                        0.00208
```

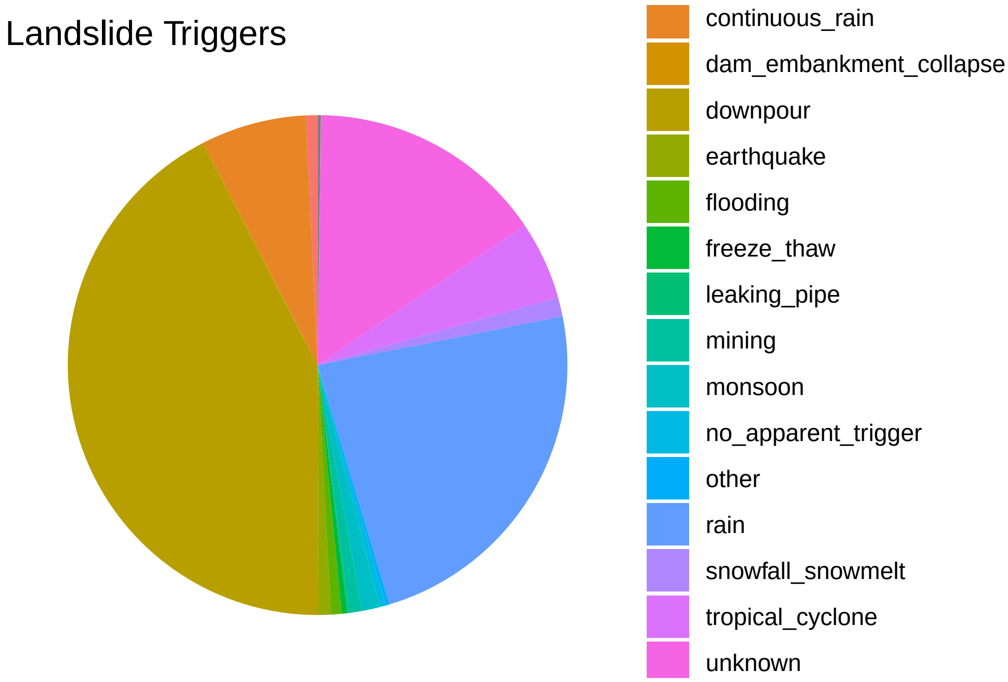
```r
# Scatter plot of longitude vs. latitude colored by landslide size
ggplot(data, aes(x = longitude, y = latitude, color = landslide_size)) +
  geom_point() +
  labs(title = "Geographic Distribution of Landslides by Size", x = "Longitude", y = "Latitu
```



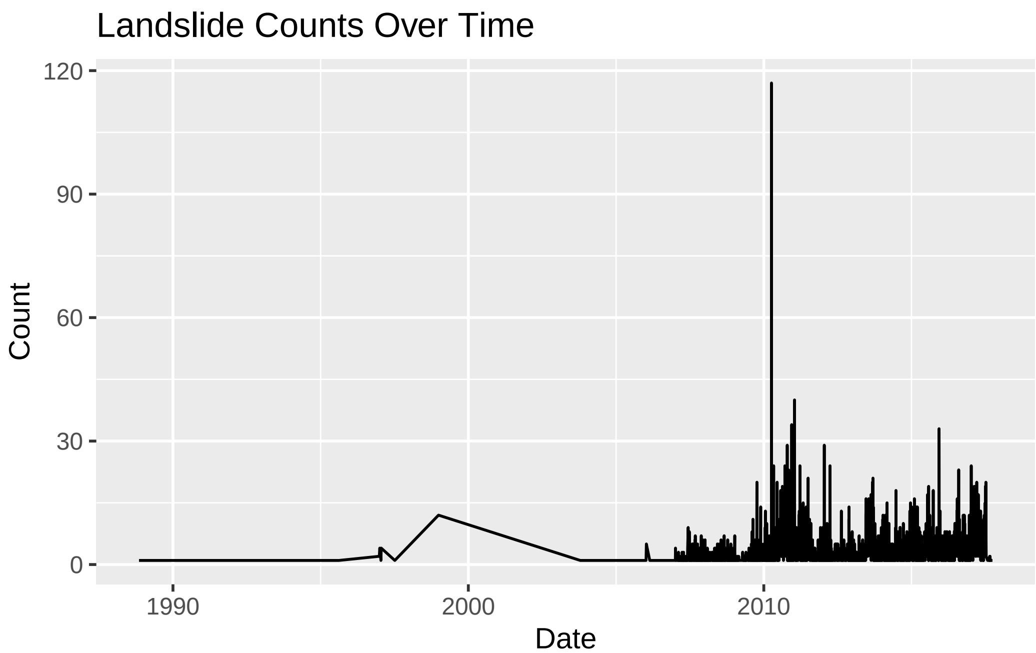

```r
# Density plot of landslide occurrence by date
ggplot(data, aes(x = date)) +
  geom_density() +
  labs(title = "Density of Landslide Occurrences by Date", x = "Date", y = "Density")
```



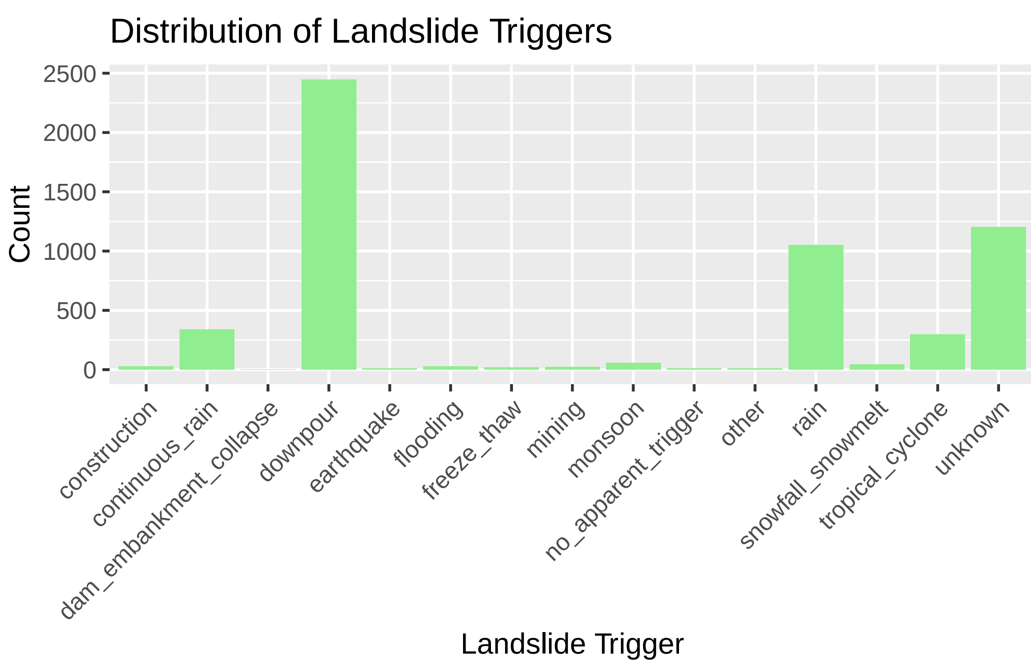

```r
# Pie chart of landslide triggers
ggplot(data, aes(x = "", fill = landslide_trigger)) +
  geom_bar(width = 1, stat = "count") +
  coord_polar("y", start = 0) +
  labs(title = "Landslide Triggers", fill = "Landslide Trigger") +
  theme_void()
```



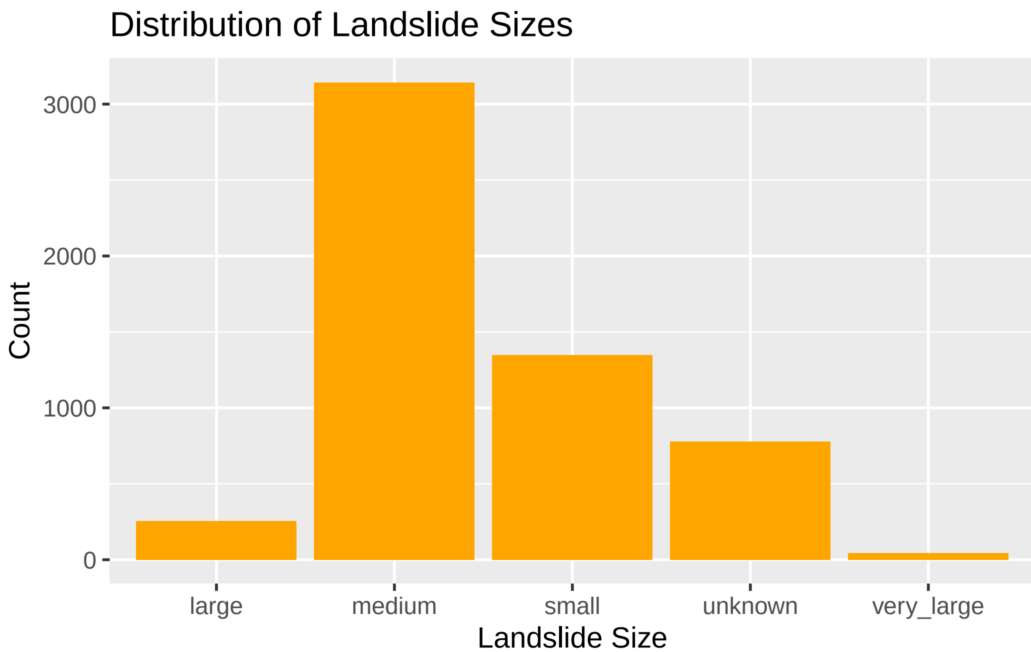

```r
# Line plot of landslide counts over time
ggplot(data, aes(x = date, group = 1)) +
  geom_line(stat = "count") +
  labs(title = "Landslide Counts Over Time", x = "Date", y = "Count")
```



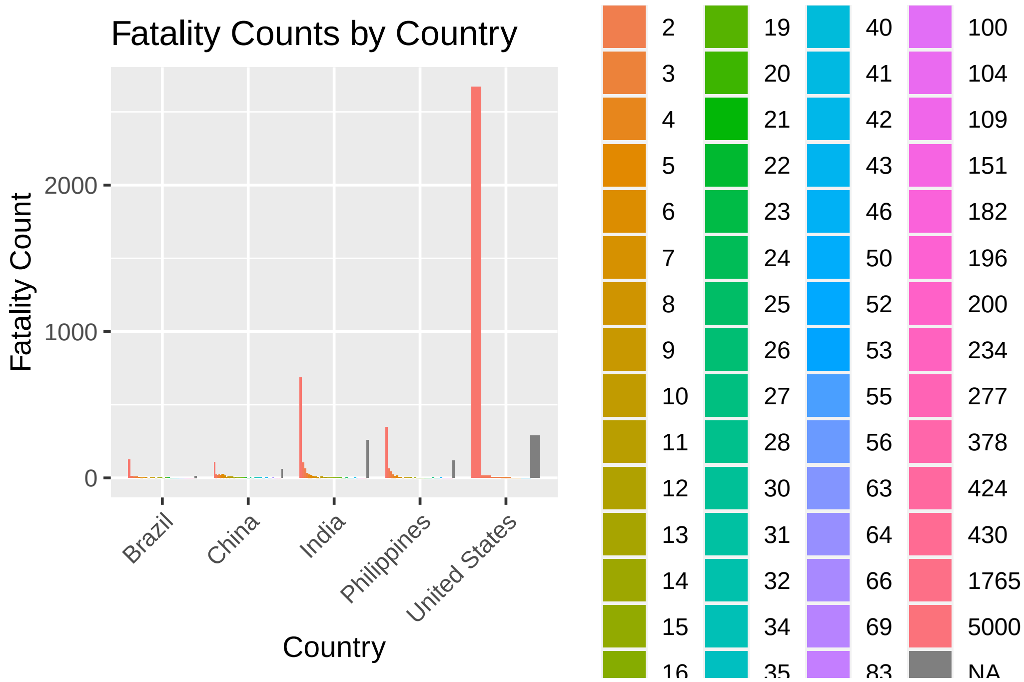
```r
sample_countries <- c("United States", "India", "China", "Brazil", "Philippines")

sample_data <- data |>
  filter(country_name %in% sample_countries)

# Bar chart of landslide trigger
ggplot(sample_data, aes(x = landslide_trigger)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Distribution of Landslide Triggers",
       x = "Landslide Trigger", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```r
# Bar chart of landslide size
ggplot(sample_data, aes(x = landslide_size)) +
  geom_bar(fill = "orange") +
  labs(title = "Distribution of Landslide Sizes",
       x = "Landslide Size", y = "Count")
```




```
# Bar chart of country-wise fatality count
ggplot(sample_data, aes(x = country_name, fill = factor(fatality_count))) +
  geom_bar(position = "dodge", width = 0.8) +
  labs(title = "Fatality Counts by Country",
       x = "Country", y = "Fatality Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Questions for reviewers

For Peer Reviewers:

Are the visualizations (graphs, plots, etc.) clear and informative? Do they effectively convey the patterns and insights you're trying to highlight? Are there any additional visualizations or exploratory analyses that you think would be beneficial to include, given the research questions? Do you have any concerns or feedback regarding the data cleaning or transformation steps performed? Are there any potential biases or limitations in the data or analysis that you think should be addressed or discussed further? Do you have any suggestions for improving the clarity or structure of the data description section?

For Project Mentor:

Given the research questions and the data available, do you think the exploratory analysis and visualizations adequately address the key aspects of the questions? Are there any specific analytical approaches or statistical techniques that you would recommend for further investigating the research questions? Do you have any feedback or suggestions regarding the handling of missing data or the normalization of variables (if applicable)? Are there any potential confounding factors or variables that should be considered in the analysis to strengthen the findings? Do you have any concerns or recommendations regarding the interpretation of the results or the validity of the conclusions drawn from the exploratory analysis? Can you provide guidance on how to effectively communicate the limitations and potential biases of the data or analysis in the final report or presentation?