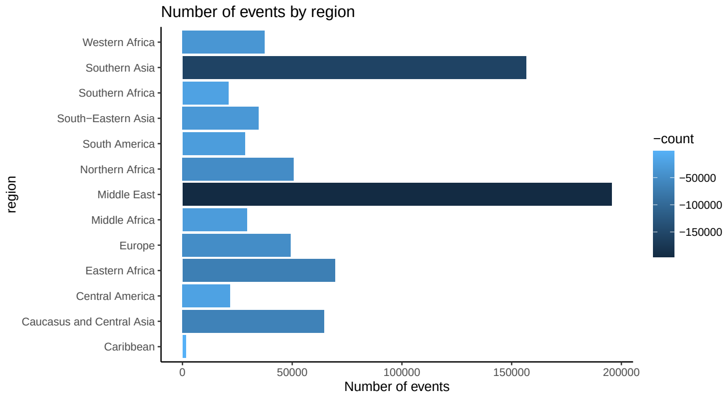
```
> library(dplyr)
> library(ggplot2)
> #library(mapview)
> library(tidyverse)
> library(ggmap)
> # Data upload
> proj_data = read.csv("acled_2020.csv")
> sample_data = read.csv("Shenoy Data Sample ACLED.csv")
> var_int = c("region", "country","admin1",
+             "admin2","admin3","location", "latitude",
+             "longitude","geo_precision","source","source_scale")
> eda = proj_data[, var_int]
> # Checking missing values
> sum(is.na(eda))
```

There are no missing values in these variables in the data.

```
> #--------------------------------------------------------------------
> # The variable region
> #--------------------------------------------------------------------
> region.t = table(eda$region)
> region.df = data.frame(region.t)
> colnames(region.df) = c("region", "freq")
> #ggplot(stg, aes( x = StudentID) ) + geom_bar()
> ggplot(data = eda, aes(y = region)) +
+   geom_bar(color = "orange", fill = "orange")
> region_plot <- eda %>% group_by(region) %>%summarise(count=n())%>%arrange(desc(count))
> region_plot %>% ggplot(aes(y=region, x = count, fill = -count)) +
+    geom_bar(stat = "identity") +
+    ggtitle("Number of events by region") +
+    scale_x_continuous(name = "Number of events") +
+   theme_classic()
```

**region:** The data is recorded for 13 regions. In terms of the number of events, Middle East region is ranked first and Caribbean region tanks last. The second rank goes to South Asia.

```
> #---------------------------------------------------------------------------
> # The variable country
> #---------------------------------------------------------------------------
> # country.t = table(eda$country)
> # country.df = data.frame(country.t)
> # colnames(country.df) = c("country", "freq")
> # country.df = country.df[order(-country.df$freq, decreasing = FALSE),]
> # row.names(country.df) <- 1:nrow(country.df)
> # country.df[1:10,]
> # country.df[139:148,]
```

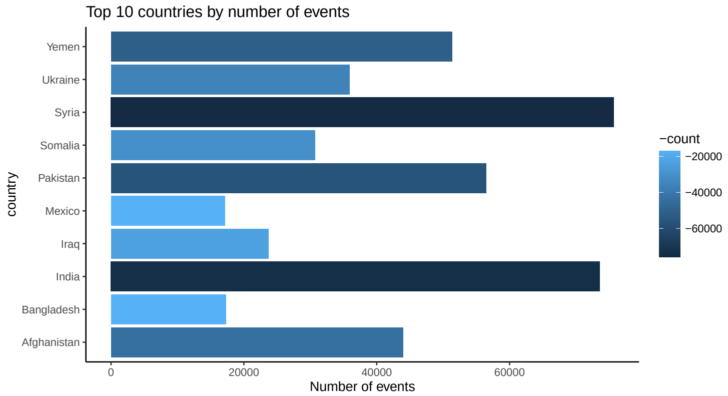


```
> country_summ <- eda %>% group_by(country) %>%summarise(count=n())%>%arrange(desc(count))
> country_summ[1:10, ]
> country_summ[1:10, ] %>% ggplot(aes(y= country, x = count, fill = -count)) +
+    geom_bar(stat = "identity") +
+    ggtitle("Top 10 countries by number of events") +
+    scale_x_continuous(name = "Number of events") +
+   theme_classic()
>
> # Summarizing
>
```

The top 6 countries in terms of number of events are Syria (75715), India (73609),Pakistan (56444), Yemen (51360), Afghanistan (43938), and Ukraine (35900). The next four countries in decreasing orders are Somalia, Iraq, Bangladesh, and Mexico. And the bottom six countries are Caribbean Netherlands, Saint Kitts and Nevis, Anguilla, Falkland Islands, Montserrat, and Virgin Islands, U.S.

The total number of events in Syria and Yemen is smaller than the total number of events in India and Pakistan. Also, the top ten list includes Afghanistan and Bangladesh. This suggests that there are other countries in Middle East which also have more events.

```
> #------------------------------------------------------------------------------
> # the variable admin 1
> #------------------------------------------------------------------------------
> # eda$admin1
> # admin1.t = table(eda$admin1)
> # admin1.df = data.frame(admin1.t)
> # names(admin1.df) = c("admin1", "freq")
> # admin1.df = admin1.df[order(admin1.df$freq, decreasing = TRUE),]
> admin1_summ <- eda %>% group_by(admin1) %>%summarise(count=n())%>%arrange(desc(count))
> admin1_summ[1:10, ]
```



```
> admin1_summ[1:10, ] %>% ggplot(aes(y= admin1, x = count, fill = -count)) +
+    geom_bar(stat = "identity") +
+    ggtitle("Top 10 provincial level administrations by number of events") +
+    scale_x_continuous(name = "Number of events") +
+   theme_classic()
```

The variable admin1 refers to the administrative unit equivalent to provincial level or similar. There are 2250 provincial level or equivalent administration where event occurred. The largest number of events occurred in Sindh (27480), and in decreasing order Donetsk (25367), Punjab (18402), Idleb(15467), Hama(14688), Jammu and Kashmir(12612). Similarly, the smallest numbers are for Wanica, Yardymli, Yauco, Zaqatala, Zardab, Znaur.

```
> #-------------------------------------------------------------------------
> # the variable admin2
> #-------------------------------------------------------------------------
> admin2_summ <- eda %>% group_by(admin2) %>%summarise(count=n())%>%arrange(desc(count))
> admin2_summ[1:10, ]
> admin2_summ[2:11, ] %>% ggplot(aes(y= admin2, x = count)) +
+    geom_bar(stat = "identity", fill = "darkgreen") +
+    ggtitle("Top 10 provincial level administrations by number of events excluding missing
+    scale_x_continuous(name = "Number of events") +
+   theme_classic()
```

The variable admin2 refers to the administrative unit equivalent to county level or similar. There are 14637 records of such administrative units where events occurred. There missing county level units for 25263 events.
**Question:** What do we do for this?

```
> #-------------------------------------------------------------------------
> # the variable admin3
> #-------------------------------------------------------------------------
> admin3_summ <- eda %>% group_by(admin3) %>%summarise(count=n())%>%arrange(desc(count))
> admin3_summ[1:10, ]
> admin3_summ[2:11, ] %>% ggplot(aes(y= admin3, x = count)) +
+    geom_bar(stat = "identity", fill = "darkgreen") +
+    ggtitle("Top 10 district level administrations by number of events excluding missing pl
+    scale_x_continuous(name = "Number of events") +
+   theme_classic()
```

The variable admin3 refers to the administrative unit equivalent to district level or similar. There are 14687 records of such administrative units where events occurred. There missing district level units for 384122 events.

```
> #---------------------------------------------------------------------------
> # the variable location
> #---------------------------------------------------------------------------
> summary(eda$location)
> location_summ <- eda %>% group_by(location) %>%summarise(count=n())%>%arrange(desc(count))
> location_summ[1:10, ] %>% ggplot(aes(y= location, x = count)) +
+    geom_bar(stat = "identity", fill = "darkorange") +
+    ggtitle("Top 10 locations by number of events") +
+    scale_x_continuous(name = "Number of events") +
+   theme_classic()
```

The variable location refers to a village or a town. Karachi, Hyderabad, and Peshawar, Jammu, and Larkana are top five locations with respect to the number of events with each more than 2000 events. There are more 30 locations with more than 1000 events and 1465 locations with more than 100 events. There are 26823 locations with only one event.

```
> # my_sf <- st_as_sf(eda[, c('latitude','longitude')], coords = c('latitude', 'longitude'),
> # ggplot(my_sf) +
> #   geom_sf()
>

> #---------------------------------------------------------------------------
> # the variable latitude
> #---------------------------------------------------------------------------
> summary(eda$latitude)
```

The variable latitude ranges from -54.92 to 69.35 with median 25.97.

```
> #---------------------------------------------------------------------------
> # the variable longitude
> #---------------------------------------------------------------------------
> summary(eda$longitude)
> #The variable latitude ranges from -117.08 to 117.51 with median 42.67.
> #---------------------------------------------------------------------------
> # the variable geo_precision
> #---------------------------------------------------------------------------
> precision_summ <- eda %>% group_by(geo_precision) %>%summarise(count=n())%>%arrange(desc(c
> precision_summ %>% ggplot(aes(y= count, x = geo_precision)) +
+    geom_bar(stat = "identity", fill = "darkgreen") +
+    ggtitle("Summary of geo_precision") +
+    scale_x_continuous(name = "geo_precision_type") +
+   theme_classic()
>
>
```

The variable geo-precision ranges from 1 to 3. Here 1 refers to the highest precision; 2 refers to the general area where an event took place (not the precise location); and 3 means a larger region is mentioned (it is the closest natural location). There are 217453 events labelled as 2 and 24284. Approximately, sixty eight percentage of events are reported with highest precision.

```
> #---------------------------------------------------------------------------
> # the variable source
> #---------------------------------------------------------------------------
> source_summ <- eda %>% group_by(source) %>%summarise(count=n())%>%arrange(desc(count))
> source_summ[1:10,] %>% ggplot(aes(y= source, x = count)) +
+    geom_bar(stat = "identity", fill = "darkgreen") +
+    ggtitle("top 10 sources") +
+    scale_x_continuous(name = "geo_precision_type") +
+   theme_classic()
> rel_freq <- source_summ$count/sum(source_summ$count)
> source_summ$rel_freq = rel_freq
> source_summ[1:10,]
```

Around 32319 sources contributed to data reporting. SOHR contributed around 4% of source for the events, and 3.6% came from undisclosed sources.

```
> #---------------------------------------------------------------------------
> # the variable source scale
>
> #---------------------------------------------------------------------------
> source_scale_summ <- eda %>% group_by(source_scale) %>%summarise(count=n())%>%arrange(desc
> source_scale_summ  %>% ggplot(aes(y= source_scale, x = count)) +
+    geom_bar(stat = "identity", fill = "darkgreen") +
+    ggtitle("summary of source scale") +
+    scale_x_continuous(name = "geo_precision_type") +
+   theme_classic()
>
```

Source scale refers to level at which source used for coding an event operate. A large number of sources operate at National level.