Exploratory Analysis of Conflict Dataset

Mohsyn Imran Malik

Exploratory Analysis

Here is our final dataset created from last week

`combo_df`

First, we will load the `tidyverse and dplyr` package to aid in our exploratory analysis

```
library(tidyverse)
library(dplyr)
```

Lets start by looking at head and tail of dataset

```
head(combo_df)
```

```
tail(combo_df)
```

We can print summary stats (mean, median and interquartile range) of our dataset and get an idea of the missing data (NA)

```
summary_df <- summary(combo_df)
```

```
summary(combo_df)
```

We can also summarize by country or year

```
summary_country <- by(combo_df, combo_df$ISO, summary)
```

```
summary_country
```

```
summary_year <- by(combo_df, combo_df$Year, summary)
summary_year
```

Note that we have significant amounts of missing data (NA) for a number of variables in the dataset.

Lets compare mean and standard deviation of continuous variables data when we remove all rows with missing data

```
c <- c(3:5, 9, 12, 15:17, 19:22)

mean <- sapply(combo_df[, c], mean, na.rm = TRUE)

mean

sd <- sapply(combo_df[, c], sd, na.rm = TRUE)

sd
```

Now lets remove rows with missing data and find new means and sd

```
remove_na <- combo_df[complete.cases(combo_df[ , c]),]

rem_mean <- sapply(remove_na[, c], mean, na.rm = TRUE)

rem_mean

 rem_sd <- sapply(remove_na[, c], sd, na.rm = TRUE)

rem_sd
```

We can do the same but this time by using multiple imputations to account for missing data

```
library(mice)
mult_imp <- mice(combo_df)
mice_est <- complete(mult_imp) |> as_tibble()

imp_mean <- sapply(mice_est[, c], mean, na.rm = TRUE)
imp_mean
imp_sd <- sapply(mice_est[, c], sd, na.rm = TRUE)
imp_sd
```

Use cbind to visualize the differences in means on method of dealing with missing data in table format

```
md_df <- cbind(mean, rem_mean, imp_mean, sd, rem_sd, imp_sd)

md_df<- as.data.frame(md_df)

md_df<- rownames_to_column(md_df, var = "variables")

md_df
```

Lets make a visual plot of contingency tables using mosaicplots.
We can look at year and binary conflict variable.

```
 mosaicplot(table(combo_df$Year, combo_df$bin_conflict), color = TRUE, xlab = "Year", ylab =
```

We can visualize the spread of data using histograms for our continuous variables #Lets look at total droughts, total earthquakes, total battle related death, gdp100, popdens, male_edu, temp, rainfall1000, maternal.mortality, infant.mortality, neonatal.mortality, and under 5 mortality

```
num_combo_df <- sapply(combo_df, as.numeric)

for (i in c(3:5, 9, 12, 15:17, 19:22)) { hist(num_combo_df[, i], main = paste("Histogram of"
```

Lets use a bar graft to examine the binary variable of conflict for all countries and years included

```
ggplot(data = combo_df) + geom_bar(mapping = aes(x = bin_conflict))
```

Lets plot gdp100 by conflict using ggplot

```
ggplot(data = combo_df) + geom_point(mapping = aes(x = gdp1000, y = bin_conflict))
```

Lets make a comparative histogram to see how bins of gdp compare in those with conflict and those without.

```
 ggplot(data = combo_df, aes(x = gdp1000, group = bin_conflict, fill = bin_conflict )) + geo
```

Lets further examine this relationship by creating a new variable which divides data into quartiles based on male education

```
combo_df$male_edu_qt <- ntile(combo_df$male_edu, 4)
ggplot(data = combo_df, aes(x = gdp1000, group = bin_conflict, fill = bin_conflict )) + geom_
```

Let us examine the trend in maternal mortality over the years and stratify into separate plots based on conflict

```
ggplot(data = combo_df, aes(x = Year, y = Maternal.Mortality)) + geom_point() + geom_smooth()
```

Lets do the same plot, but use our dataset with multiply imputed missing variables

```
ggplot(data = mice_est, aes(x = Year, y = Maternal.Mortality)) + geom_point() + geom_smooth()
```