CHL8010: Statistical Programming and Computation in Health Data

2024-09-30

Observations from Canada should look like this...

```
Final_Canada <- Final_data %>%
  dplyr::filter(country_name == "Canada")

head(Final_Canada, 10)
```

```
   country_name ISO           region  gdp1000 OECD OECD2023  popdens    urban
1        Canada CAN Northern America 24.27100    1        1 66.19704 56.14335
2        Canada CAN Northern America 23.82206    1        1 66.45361 56.40270
3        Canada CAN Northern America 24.25534    1        1 66.71112 56.67093
4        Canada CAN Northern America 28.30046    1        1 66.96384 56.94365
5        Canada CAN Northern America 32.14368    1        1 67.21715 57.20020
6        Canada CAN Northern America 36.38251    1        1 67.47283 57.41671
7        Canada CAN Northern America 40.50406    1        1 67.73674 57.59143
8        Canada CAN Northern America 44.65990    1        1 67.99444 57.75691
9        Canada CAN Northern America 46.71051    1        1 68.25765 57.97905
10       Canada CAN Northern America 40.87631    1        1 68.53354 58.24228
     agedep male_edu     temp rainfall1000 Year Totdeath Conflict MatMor NeoMor
1  46.34463 12.30281 5.486244    0.9971559 2000       11        0      9    3.8
2  45.89632 12.35258 6.469105    0.8644873 2001       23        0     10    3.8
3  45.46660 12.40182 5.979147    0.9460938 2002        1        0     10    3.9
4  45.07468 12.45053 5.416964    1.0189234 2003        0        0     10    3.9
5  44.67374 12.49870 5.556961    1.0008237 2004        0        0     10    3.9
6  44.26641 12.54635 6.187472    1.0367199 2005        0        0     11    3.9
7  43.96370 12.59349 6.895084    1.0917386 2006        0        0     11    3.9
8  43.83612 12.64015 5.900051    1.0134091 2007        0        0     11    3.8
9  43.85426 12.68634 5.650118    1.0693435 2008        0        0     12    3.8
```

```
10 43.94937 12.73207 5.398867    0.9928497 2009        0        0     12    3.8
   InfMor Und5Mor drought earthquake
1     5.3     6.2       0          0
2     5.3     6.2       0          0
3     5.3     6.2       0          0
4     5.3     6.2       0          0
5     5.3     6.1       0          0
6     5.2     6.1       0          0
7     5.2     6.0       0          0
8     5.1     6.0       0          0
9     5.1     5.9       0          0
10    5.0     5.8       0          0
```

Observations from Ecuador should look like this...

```
Final_Equador <- Final_data %>%
  dplyr::filter(country_name == "Ecuador")
head(Final_Equador, 10)
```

```
   country_name ISO                          region  gdp1000 OECD OECD2023
1       Ecuador ECU Latin America and the Caribbean 1.451531    0        0
2       Ecuador ECU Latin America and the Caribbean 1.904814    0        0
3       Ecuador ECU Latin America and the Caribbean 2.184209    0        0
4       Ecuador ECU Latin America and the Caribbean 2.438344    0        0
5       Ecuador ECU Latin America and the Caribbean 2.703566    0        0
6       Ecuador ECU Latin America and the Caribbean 3.014310    0        0
7       Ecuador ECU Latin America and the Caribbean 3.340841    0        0
8       Ecuador ECU Latin America and the Caribbean 3.579032    0        0
9       Ecuador ECU Latin America and the Caribbean 4.260433    0        0
10      Ecuador ECU Latin America and the Caribbean 4.240703    0        0
    popdens    urban   agedep male_edu     temp rainfall1000 Year Totdeath
1  23.27432 36.19963 67.44216 7.738627 19.54855    1.4201653 2000        0
2  23.39372 36.67994 66.57356 7.843942 19.66622    1.1667746 2001        0
3  23.52087 37.08903 65.65488 7.949449 20.24695    1.4577981 2002        2
4  23.58358 37.23792 64.71472 8.055240 20.05016    1.5781807 2003        0
5  38.43743 37.39268 63.78049 8.161433 20.10136    1.0683450 2004       26
6  38.55361 37.36968 62.86530 8.268176 19.88163    0.8555447 2005        0
7  38.65018 37.47567 61.97042 8.375587 20.07087    1.1114502 2006        0
8  38.76505 37.68172 61.11422 8.483729 19.49536    1.0899082 2007        0
9  38.83977 37.67445 60.31015 8.592603 19.85711    1.6184816 2008        0
10 38.92613 37.39437 59.55262 8.702180 20.39298    1.0870796 2009       25
   Conflict MatMor NeoMor InfMor Und5Mor drought earthquake
```

```
1         0    122   14.1   24.7    29.5        0          0
2         0    117   13.4   23.4    28.0        0          0
3         0    110   12.7   22.4    26.6        0          0
4         0    100   12.1   21.5    25.4        0          0
5         1     94   11.6   20.7    24.4        0          0
6         0     94   11.1   19.9    23.5        0          0
7         0     90   10.6   19.2    22.6        0          0
8         0     85   10.2   18.5    21.7        0          0
9         0     82    9.7   17.7    20.8        0          0
10        1     80    9.3   17.0    19.9        1          0
```

Exploratory data analysis

Use the rest of the class time to explore the final data that will be used for analysis starting next week. At the end of the class, write a summary of your findings and push your **Quarto document (pdf)** to your repo.

```r
head(Final_data)
```

```
  country_name ISO        region   gdp1000 OECD OECD2023  popdens    urban
1  Afghanistan AFG Southern Asia        NA    0        0 14.13654 16.25324
2  Afghanistan AFG Southern Asia        NA    0        0 14.23156 16.25661
3  Afghanistan AFG Southern Asia 0.1835328    0        0 14.32270 16.42654
4  Afghanistan AFG Southern Asia 0.2004626    0        0 14.40691 16.60701
5  Afghanistan AFG Southern Asia 0.2216576    0        0 15.21947 16.71367
6  Afghanistan AFG Southern Asia 0.2550551    0        0 15.33619 16.85096
    agedep male_edu     temp rainfall1000 Year Totdeath Conflict MatMor NeoMor
1 108.3466 2.762086 12.69959    0.2763704 2000     5065        1   1450   60.9
2 108.9899 2.856936 12.85570    0.2793079 2001     5394        1   1390   59.7
3 109.3472 2.954241 12.71081    0.3805710 2002     5553        1   1300   58.5
4 109.4475 3.054121 12.16592    0.4288939 2003     1157        1   1240   57.2
5 109.2868 3.156706 13.04643    0.3754336 2004      944        1   1180   55.9
6 107.9646 3.262133 12.23141    0.4415680 2005      817        1   1140   54.6
  InfMor Und5Mor drought earthquake
1   90.5   129.2       1          0
2   87.9   125.2       0          2
3   85.3   121.1       0          3
4   82.7   116.9       0          1
5   80.0   112.6       0          1
6   77.3   108.4       0          2
```

```
tail(Final_data)
```

```
     country_name ISO             region  gdp1000 OECD OECD2023  popdens
3715     Zimbabwe ZWE Sub-Saharan Africa 1.407034    0        0 26.52884
3716     Zimbabwe ZWE Sub-Saharan Africa 1.410329    0        0 26.54454
3717     Zimbabwe ZWE Sub-Saharan Africa 1.421788    0        0 26.53811
3718     Zimbabwe ZWE Sub-Saharan Africa 1.192107    0        0 26.49281
3719     Zimbabwe ZWE Sub-Saharan Africa 2.269177    0        0 26.47943
3720     Zimbabwe ZWE Sub-Saharan Africa 1.421869    0        0 26.46341
        urban   agedep male_edu     temp rainfall1000 Year Totdeath Conflict
3715 24.40427 85.87550 8.679591 20.87651    0.6777257 2014        0        0
3716 24.75233 85.08337 8.785078 21.45470    0.4490721 2015        0        0
3717 25.02842 84.11222 8.889947 21.39290    0.4939246 2016        0        0
3718 25.29333 83.10129 8.994252 20.85962    0.9533149 2017        0        0
3719 25.53759 82.12335 9.098048 20.86041    0.9535655 2018        0        0
3720 25.70572 81.20786 9.201384 20.86120    0.9538138 2019        4        0
     MatMor NeoMor InfMor Und5Mor drought earthquake
3715    494   28.2   42.9    62.7       0          0
3716    480   27.8   42.1    61.3       0          0
3717    468   27.4   40.8    58.7       0          0
3718    458   27.0   39.9    57.0       1          0
3719     NA   26.6   38.8    54.8       0          0
3720     NA   26.2   38.1    54.2       0          0
```

```
dim(Final_data)
```

```
[1] 3720   21
```

```
str(Final_data)
```

```
'data.frame':	3720 obs. of  21 variables:
 $ country_name: chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
 $ ISO         : chr  "AFG" "AFG" "AFG" "AFG" ...
 $ region      : chr  "Southern Asia" "Southern Asia" "Southern Asia" "Southern Asia" ...
 $ gdp1000     : num  NA NA 0.184 0.2 0.222 ...
 $ OECD        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OECD2023    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ popdens     : num  14.1 14.2 14.3 14.4 15.2 ...
 $ urban       : num  16.3 16.3 16.4 16.6 16.7 ...
 $ agedep      : num  108 109 109 109 109 ...
 $ male_edu    : num  2.76 2.86 2.95 3.05 3.16 ...
 $ temp        : num  12.7 12.9 12.7 12.2 13 ...
 $ rainfall1000: num  0.276 0.279 0.381 0.429 0.375 ...
 $ Year        : int  2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 ...
 $ Totdeath    : int  5065 5394 5553 1157 944 817 1711 4982 7020 5660 ...
 $ Conflict    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ MatMor      : int  1450 1390 1300 1240 1180 1140 1120 1090 1030 993 ...
 $ NeoMor      : num  60.9 59.7 58.5 57.2 55.9 54.6 53.2 51.7 50.3 48.9 ...
 $ InfMor      : num  90.5 87.9 85.3 82.7 80 77.3 74.6 71.9 69.2 66.7 ...
 $ Und5Mor     : num  129 125 121 117 113 ...
 $ drought     : int  1 0 0 0 0 0 1 0 1 0 ...
 $ earthquake  : int  0 2 3 1 1 2 1 0 0 1 ...
```

3720 rows by 21 columns Data types include: 3 character, 8 integer, and 10 numerical

```
# Data shows 186 countries each with 20 rows
head(Final_data %>% count(ISO),10)
```

```
   ISO  n
1  AFG 20
2  AGO 20
3  ALB 20
4  AND 20
5  ARE 20
6  ARG 20
7  ARM 20
8  ATG 20
9  AUS 20
10 AUT 20
```

```
Final_data %>% count(ISO) %>% count(n)
```

```
Storing counts in `nn`, as `n` already present in input
i Use `name = "new_name"` to pick a new name.
```

```
   n  nn
1 20 186
```

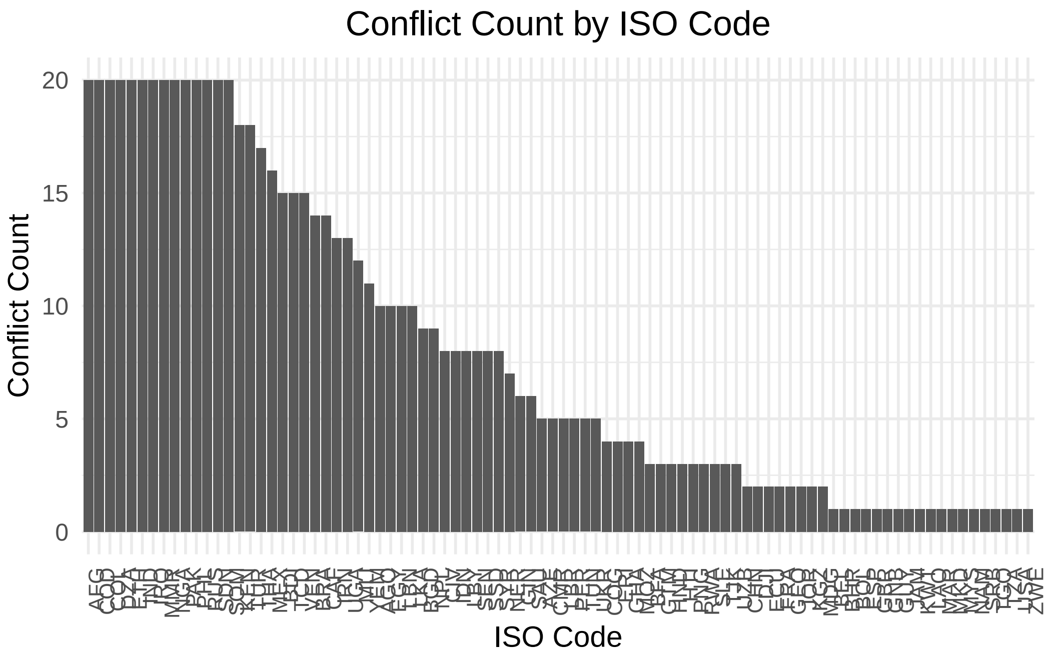
```r
#Plotting conflicts by ISO Code (only for countries with minimum 1 conflict)
conflict_data <- Final_data %>%
  group_by(ISO) %>%
  summarise(total_conflicts = sum(Conflict, na.rm = TRUE)) %>%
  filter(total_conflicts > 0) %>%
  arrange(desc(total_conflicts))
conflict_data
```

```
# A tibble: 88 x 2
   ISO   total_conflicts
   <chr>           <int>
 1 AFG                20
 2 COD                20
 3 COL                20
 4 DZA                20
 5 ETH                20
 6 IND                20
 7 IRQ                20
 8 MMR                20
 9 NGA                20
10 PAK                20
# i 78 more rows
```

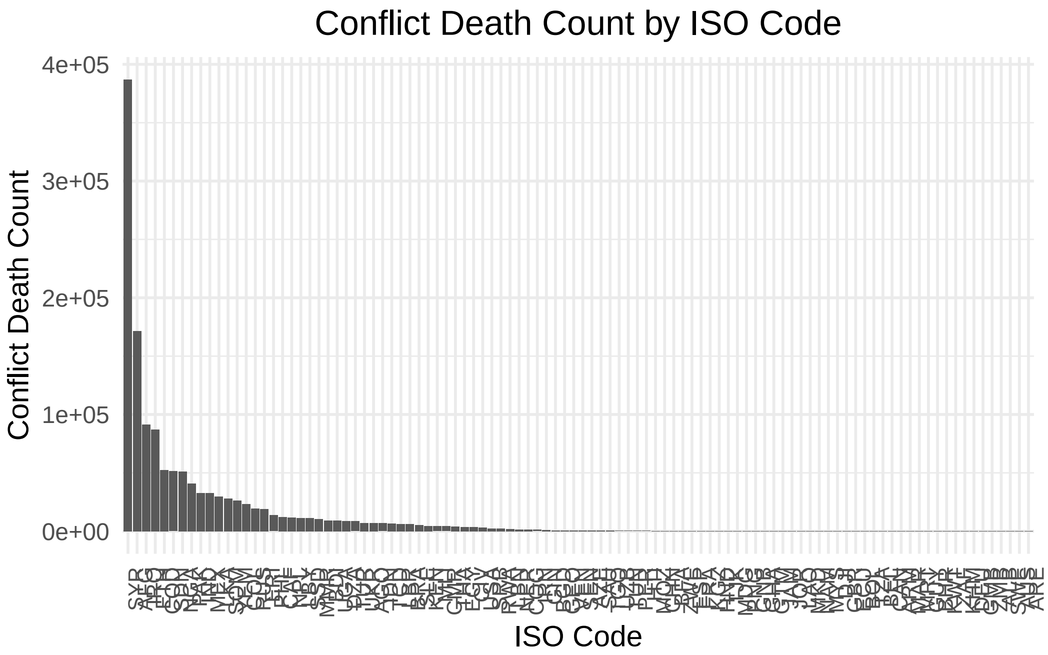
```r
#Plotting deaths by ISO Code (only for countries with minimum 1 conflict related death)
death_data <- Final_data %>%
  group_by(ISO) %>%
  summarise(total_deaths = sum(Totdeath, na.rm = TRUE)) %>%
  filter(total_deaths > 0) %>%
  arrange(desc(total_deaths))
death_data
```

```
# A tibble: 100 x 2
   ISO   total_deaths
   <chr>        <int>
 1 SYR         386891
 2 AFG         171391
 3 IRQ          91429
 4 ETH          87066
 5 COD          52492
 6 SDN          51355
 7 NGA          51114
 8 PAK          40789
 9 IND          32704
10 MEX          32686
# i 90 more rows
```

```
# Create the plot for conflict
ggplot(conflict_data, aes(x = reorder(ISO, -total_conflicts), y = total_conflicts)) +
  geom_bar(stat = "identity") +
  labs(title = "Conflict Count by ISO Code", x = "ISO Code", y = "Conflict Count") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, size = 8),
    plot.title = element_text(hjust = 0.5)
  )
```



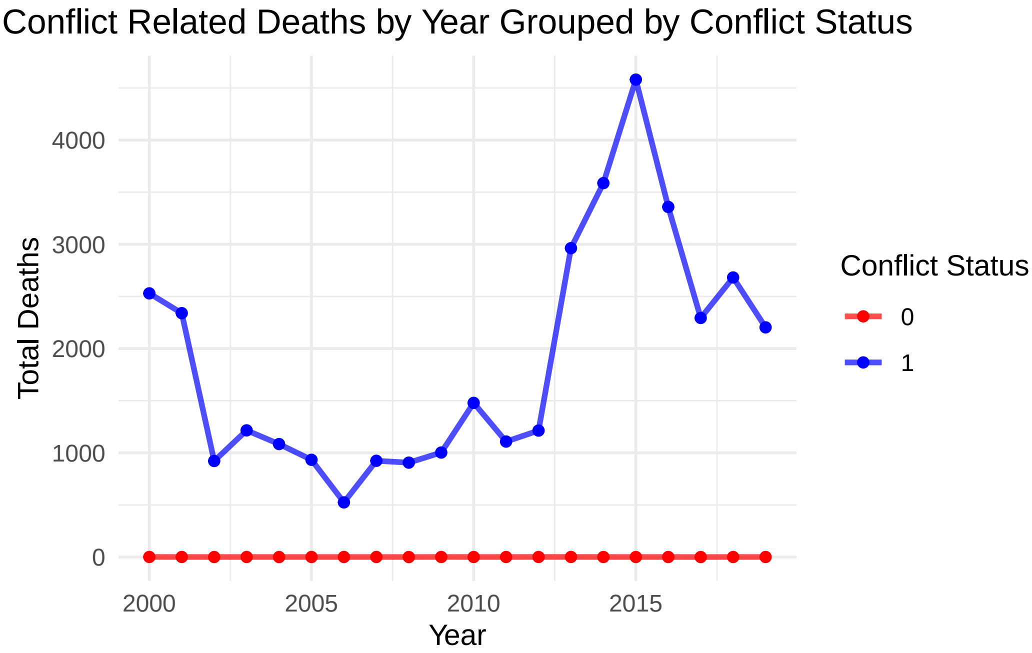
```
# Create the plot
ggplot(death_data, aes(x = reorder(ISO, -total_deaths), y = total_deaths)) +
  geom_bar(stat = "identity") +
  labs(title = "Conflict Death Count by ISO Code", x = "ISO Code",
       y = "Conflict Death Count") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 90, hjust = 1, size = 8),
    plot.title = element_text(hjust = 0.5))
```



```
conflict_summary <- Final_data %>%
  group_by(Year, Conflict) %>%
  summarise(mean = mean(Totdeath, na.rm = TRUE), .groups = 'drop')

ggplot(conflict_summary, aes(x = Year, y = mean, color = as.factor(Conflict),
                             group = Conflict)) +
  geom_line(alpha = 0.7, size = 1) +
  geom_point() +
  labs(title = "Conflict Related Deaths by Year Grouped by Conflict Status",
       x = "Year",
       y = "Total Deaths",
       color = "Conflict Status") +
  theme_minimal() +
  scale_color_manual(values = c("0" = "red", "1" = "blue"))  +
  theme(plot.title = element_text(hjust = 0.5))
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



```
#Analyze missingness
missingness <- sum(is.na(Final_data))
missingness
```

```
[1] 648
```

648 total missing values

```
missingness_column <- colSums(is.na(Final_data))
missingness_column
```

```
country_name          ISO       region      gdp1000         OECD     OECD2023 
           0            0            0           62            0            0 
     popdens        urban       agedep     male_edu         temp rainfall1000 
          20           20            0           20           20           20 
        Year     Totdeath     Conflict       MatMor       NeoMor       InfMor 
           0            0            0          426           20           20 
     Und5Mor      drought   earthquake 
          20            0            0 
```

```
missing_percent <- round(missingness_column/ nrow(Final_data) * 100,3)
missing_percent
```

```
country_name          ISO       region      gdp1000         OECD     OECD2023 
       0.000        0.000        0.000        1.667        0.000        0.000 
     popdens        urban       agedep     male_edu         temp rainfall1000 
       0.538        0.538        0.000        0.538        0.538        0.538 
        Year     Totdeath     Conflict       MatMor       NeoMor       InfMor 
       0.000        0.000        0.000       11.452        0.538        0.538 
     Und5Mor      drought   earthquake 
       0.538        0.000        0.000 
```

Maternal mortality had the most missingness of the outcome variables (N = 426 or 11.45%)

```
#Correlation of Outcome variables
round(cor(Final_data[, c("MatMor", "NeoMor", "InfMor", "Und5Mor")],
          use = "complete.obs"),4)
```

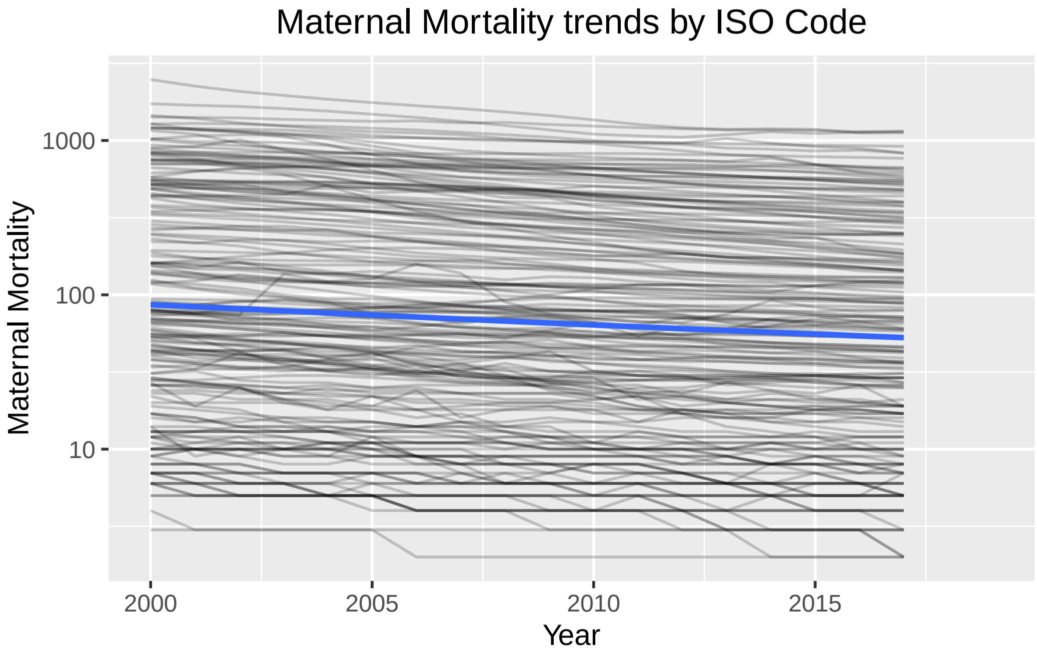
```
        MatMor NeoMor InfMor Und5Mor
MatMor  1.0000 0.8355 0.8786  0.8995
NeoMor  0.8355 1.0000 0.9591  0.9279
InfMor  0.8786 0.9591 1.0000  0.9861
Und5Mor 0.8995 0.9279 0.9861  1.0000
```

```
#Longitudinal Maternal Mortality trends by ISO Code
ggplot(Final_data, aes(Year, MatMor)) +
  geom_line(aes(group = ISO), alpha = 1/5) +
  labs(title = "Maternal Mortality trends by ISO Code", x = "Year",
       y = "Maternal Mortality") +
  scale_y_log10() +
  geom_smooth(se = FALSE) +
  theme(plot.title = element_text(hjust = 0.5))
```

```
`geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

```
Warning: Removed 426 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 426 rows containing missing values (`geom_line()`).
```


Maternal Mortality trends by ISO Code
Maternal Mortality
1000
100
10
2000
2005
2010
2015
Year