EDA_Child_Mortality_Rates_National_03

Group K

2025-09-09

```
## [1] "C:/Users/Caitlin/Documents/GitHub/BIN381-Project/scripts"
```

```
## character(0)
```

Load data and show the structure

```
cmr_df <- read_csv(here("raw_data", "child-mortality-rates_national_zaf.csv"))
```

```
## Rows: 41 Columns: 29
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (17): ISO3, DataId, Indicator, Value, Precision, DHS_CountryCode, Countr...
## dbl (10): IndicatorOrder, CharacteristicId, CharacteristicOrder, IsTotal, Is...
## lgl  (2): RegionId, LevelRank
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

#Dislay the first and last 10 rows

```
head(cmr_df, 5)
```

```
## # A tibble: 5 x 29
##   ISO3   DataId Indicator Value Precision DHS_CountryCode CountryName SurveyYear
##   <chr>  <chr>  <chr>     <chr> <chr>     <chr>           <chr>       <chr>
## 1 #coun~ #meta~ #indicat~ #ind~ #indicat~ <NA>            #country+n~ #date+year
## 2 ZAF    85995  Neonatal~ 20    0         ZA              South Afri~ 1998
## 3 ZAF    794581 Postneon~ 26    0         ZA              South Afri~ 1998
## 4 ZAF    785930 Infant m~ 45    0         ZA              South Afri~ 1998
## 5 ZAF    56239  Child mo~ 15    0         ZA              South Afri~ 1998
## # i 21 more variables: SurveyId <chr>, IndicatorId <chr>, IndicatorOrder <dbl>,
## #   IndicatorType <chr>, CharacteristicId <dbl>, CharacteristicOrder <dbl>,
## #   CharacteristicCategory <chr>, CharacteristicLabel <chr>,
## #   ByVariableId <chr>, ByVariableLabel <chr>, IsTotal <dbl>,
## #   IsPreferred <dbl>, SDRID <chr>, RegionId <lgl>, SurveyYearLabel <dbl>,
## #   SurveyType <chr>, DenominatorWeighted <dbl>, DenominatorUnweighted <dbl>,
## #   CILow <dbl>, CIHigh <dbl>, LevelRank <lgl>
```

```
#tail(cmr_df, 10)
```

#Dimenison of the data

```
dim(cmr_df)
```

```
## [1] 41 29
```

#Provide a summary of the datset

```
summary(cmr_df)
```

```
##      ISO3               DataId           Indicator            Value
##  Length:41          Length:41          Length:41          Length:41
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   Precision         DHS_CountryCode    CountryName         SurveyYear
##  Length:41          Length:41          Length:41          Length:41
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    SurveyId          IndicatorId        IndicatorOrder     IndicatorType
##  Length:41          Length:41          Min.   :63166010   Length:41
##  Class :character   Class :character   1st Qu.:63196020   Class :character
##  Mode  :character   Mode  :character   Median :63206030   Mode  :character
##                                        Mean   :63203530
##                                        3rd Qu.:63213540
##                                        Max.   :63236050
##                                        NA's   :1
##  CharacteristicId CharacteristicOrder CharacteristicCategory
##  Min.   : 1000    Min.   :    0       Length:41
##  1st Qu.: 1000    1st Qu.:    0       Class :character
##  Median : 5500    Median : 5000       Mode  :character
##  Mean   : 6250    Mean   :22500
##  3rd Qu.:10750    3rd Qu.:27500
##  Max.   :13000    Max.   :80000
##  NA's   :1        NA's   :1
##  CharacteristicLabel ByVariableId       ByVariableLabel       IsTotal
##  Length:41           Length:41          Length:41          Min.   :1
##  Class :character    Class :character   Class :character   1st Qu.:1
##  Mode  :character    Mode  :character   Mode  :character   Median :1
##                                                            Mean   :1
##                                                            3rd Qu.:1
##                                                            Max.   :1
##                                                            NA's   :1
```

```
##   IsPreferred      SDRID           RegionId       SurveyYearLabel
##  Min.   :0.00   Length:41          Mode:logical   Min.   :1998
##  1st Qu.:0.75   Class :character   NA's:41        1st Qu.:1998
##  Median :1.00   Mode  :character                  Median :2007
##  Mean   :0.75                                     Mean   :2007
##  3rd Qu.:1.00                                     3rd Qu.:2016
##  Max.   :1.00                                     Max.   :2016
##  NA's   :1                                        NA's   :1
##   SurveyType        DenominatorWeighted DenominatorUnweighted     CILow
##  Length:41          Min.   :3577        Min.   :3563          Min.   : 4.00
##  Class :character   1st Qu.:3577        1st Qu.:3563          1st Qu.:12.25
##  Mode  :character   Median :4348        Median :4373          Median :18.50
##                     Mean   :4348        Mean   :4373          Mean   :22.77
##                     3rd Qu.:5119        3rd Qu.:5183          3rd Qu.:34.50
##                     Max.   :5119        Max.   :5183          Max.   :50.00
##                     NA's   :37          NA's   :37            NA's   :11
##      CIHigh      LevelRank
##  Min.   :10.00   Mode:logical
##  1st Qu.:20.75   NA's:41
##  Median :30.00
##  Mean   :35.20
##  3rd Qu.:51.00
##  Max.   :68.00
##  NA's   :11
```

#Convert Value to a numeric

```
cmr_df$Value <- as.numeric(cmr_df$Value)
```

```
## Warning: NAs introduced by coercion
```

#Find duplicated values

```
sum(duplicated(cmr_df))
```

```
## [1] 0
```

#Inspect columns for missing/empty data as values and pecentages respectively

```
colSums(is.na(cmr_df))
```

```
##                   ISO3                 DataId              Indicator
##                      0                      0                      0
##                  Value              Precision        DHS_CountryCode
##                      1                      0                      1
##            CountryName             SurveyYear               SurveyId
##                      0                      0                      0
##            IndicatorId        IndicatorOrder          IndicatorType
##                      0                      1                      1
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                      1                      1                      1
```

```
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                      1                      0                     20
##                IsTotal            IsPreferred                 SDRID
##                      1                      1                     1
##               RegionId        SurveyYearLabel             SurveyType
##                     41                      1                     1
##    DenominatorWeighted  DenominatorUnweighted                 CILow
##                     37                     37                    11
##                 CIHigh             LevelRank
##                     11                     41
```

```
round(colMeans(is.na(cmr_df)) * 100, 2)
```

```
##                   ISO3                 DataId              Indicator
##                   0.00                   0.00                   0.00
##                  Value              Precision        DHS_CountryCode
##                   2.44                   0.00                   2.44
##            CountryName             SurveyYear               SurveyId
##                   0.00                   0.00                   0.00
##            IndicatorId         IndicatorOrder          IndicatorType
##                   0.00                   2.44                   2.44
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                   2.44                   2.44                   2.44
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                   2.44                   0.00                  48.78
##                IsTotal            IsPreferred                 SDRID
##                   2.44                   2.44                   2.44
##               RegionId        SurveyYearLabel             SurveyType
##                 100.00                   2.44                   2.44
##    DenominatorWeighted  DenominatorUnweighted                 CILow
##                  90.24                  90.24                  26.83
##                 CIHigh             LevelRank
##                  26.83                 100.00
```

Check for unique values

```
sapply(cmr_df, function(x) length(unique(x)))
```

```
##                   ISO3                 DataId              Indicator
##                      2                     41                     16
##                  Value              Precision        DHS_CountryCode
##                     28                      2                      2
##            CountryName             SurveyYear               SurveyId
##                      2                      3                      3
##            IndicatorId         IndicatorOrder          IndicatorType
##                     16                     16                      4
##       CharacteristicId    CharacteristicOrder CharacteristicCategory
##                      4                      4                      4
##    CharacteristicLabel           ByVariableId        ByVariableLabel
##                      4                      4                      4
```

```
##                 IsTotal            IsPreferred                  SDRID
##                       2                      3                     16
##              RegionId       SurveyYearLabel             SurveyType
##                       1                      3                      2
##    DenominatorWeighted  DenominatorUnweighted                  CILow
##                       3                      3                     19
##                CIHigh             LevelRank
##                    16                      1
```

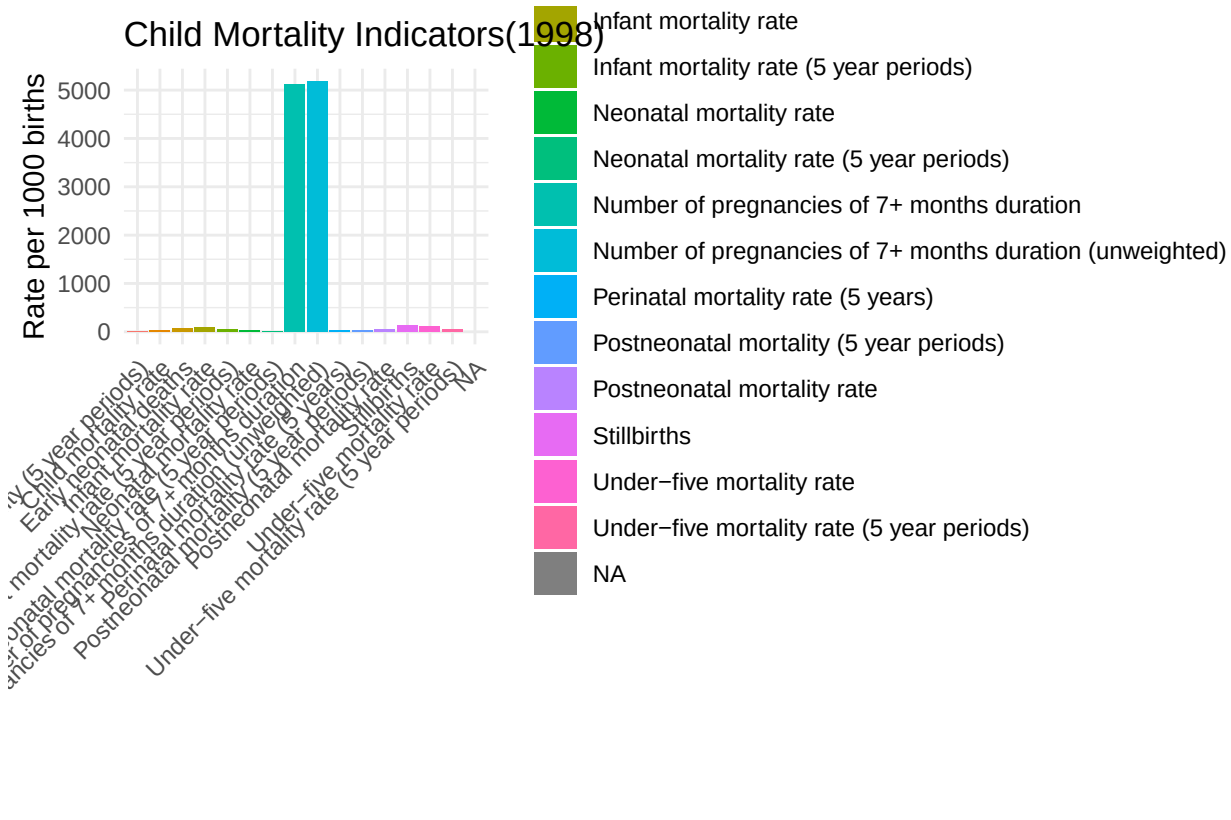
#Plots

```
cmr_df$Value <- as.numeric(cmr_df$Value)
cmr_df$SurveyYear <- as.numeric(cmr_df$SurveyYear)
```

```
## Warning: NAs introduced by coercion
```

```
# Example: Bar plot of mortality indicators for a specific year
ggplot(cmr_df[cmr_df$SurveyYear == 1998, ], aes(x = Indicator, y = Value, fill = Indicator)) +
  geom_bar(stat = "identity") +
  theme_minimal() +
  labs(title = "Child Mortality Indicators(1998)",
       x = "Indicator",
       y = "Rate per 1000 births") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

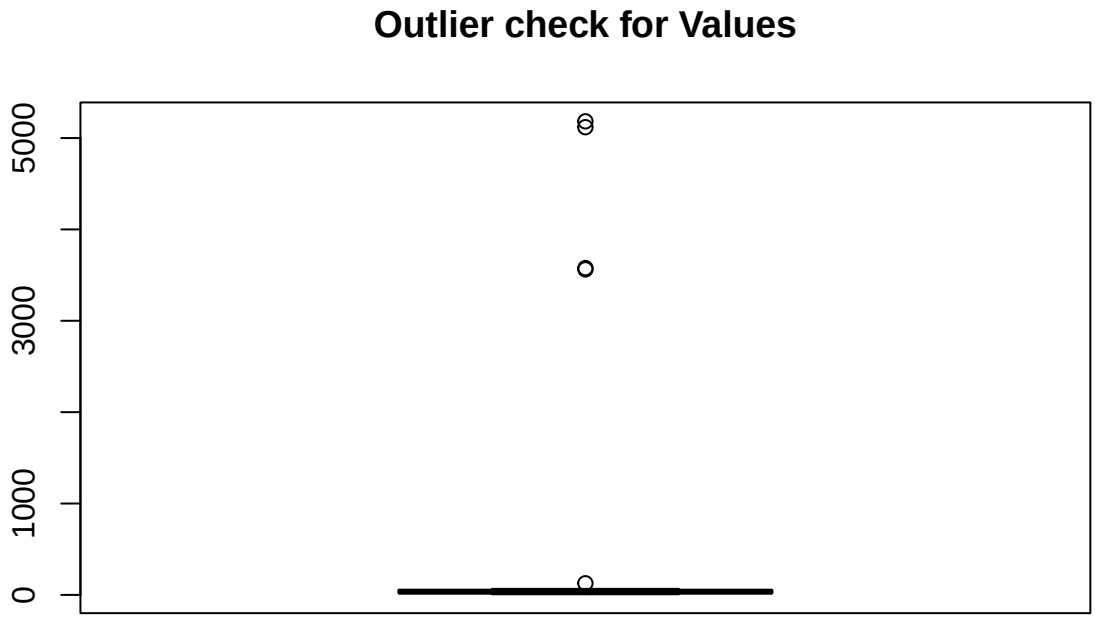
```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
ggsave("../outputs/visuals/cmr_bargraph.png", width = 6, height = 4)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```

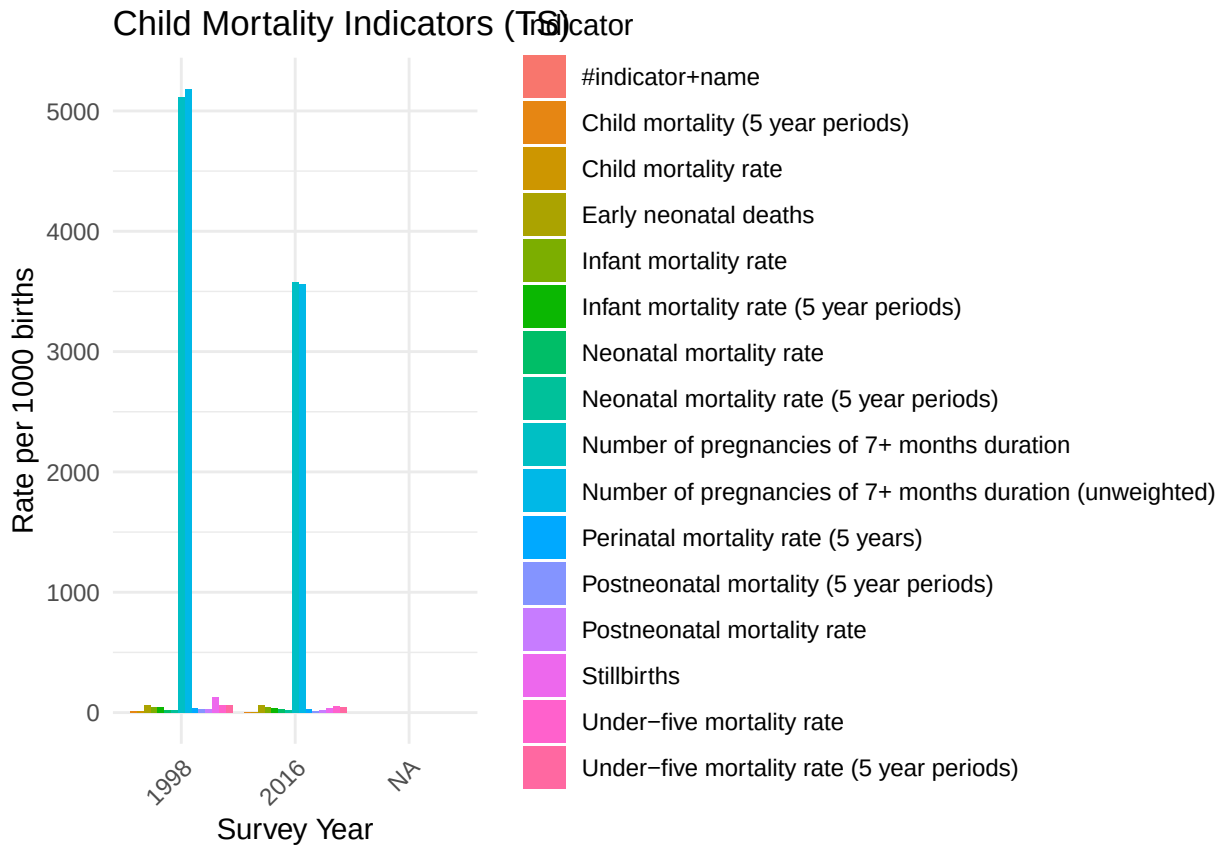
```
#Outliers
boxplot(cmr_df$Value, main = "Outlier check for Values")
```




```
cmr_df$Value <- as.numeric(cmr_df$Value)
cmr_df$SurveyYear <- as.numeric(cmr_df$SurveyYear)

ggplot(cmr_df, aes(x = factor(SurveyYear), y = Value, fill = Indicator)) +
  geom_bar(stat = "identity", position = "dodge") +
  theme_minimal() +
  labs(title = "Child Mortality Indicators (TS)",
       x = "Survey Year",
       y = "Rate per 1000 births",
       fill = "Indicator") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```




```
ggsave("../outputs/visuals/cmr_timeseriesplot.png", width = 6, height = 4)
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).
```