prepare_data

2026-02-02

This script reads data and creates structural variables needed for time series modeling and to account for trend and seasonality.

```r
library(readxl)
```

```r
raw_data <- read_excel("../data/Toddler/descriptive_toddler_Table4b.xlsx")

df <- raw_data[, !names(raw_data) %in% c("total_elig", "count")]
str(df)
```

```
## tibble [120 x 3] (S3: tbl_df/tbl/data.frame)
##  $ month_turn_2: POSIXct[1:120], format: "2015-04-01" "2015-05-01" ...
##  $ rate        : num [1:120] 70.1 70.2 70.5 69.6 70.3 ...
##  $ period      : chr [1:120] "pre" "pre" "pre" "pre" ...
```

Adding month_index for modeling the trend.

```r
df$month_index <- seq_len(nrow(df))
head(df)
```

```
## # A tibble: 6 x 4
##   month_turn_2         rate period month_index
##   <dttm>              <dbl> <chr>        <int>
## 1 2015-04-01 00:00:00  70.1 pre              1
## 2 2015-05-01 00:00:00  70.2 pre              2
## 3 2015-06-01 00:00:00  70.5 pre              3
## 4 2015-07-01 00:00:00  69.6 pre              4
## 5 2015-08-01 00:00:00  70.3 pre              5
## 6 2015-09-01 00:00:00  70.3 pre              6
```

Adding calendar month to model seasonality if needed

```r
df$calendar_month <- as.integer(format(df$month_turn_2, "%m"))
df$calendar_month <- factor(
  df$calendar_month,
  levels = 1:12,
  labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
             "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")
  )

str(df)
```

```
## tibble [120 x 5] (S3: tbl_df/tbl/data.frame)
##  $ month_turn_2  : POSIXct[1:120], format: "2015-04-01" "2015-05-01" ...
##  $ rate          : num [1:120] 70.1 70.2 70.5 69.6 70.3 ...
##  $ period        : chr [1:120] "pre" "pre" "pre" "pre" ...
##  $ month_index   : int [1:120] 1 2 3 4 5 6 7 8 9 10 ...
##  $ calendar_month: Factor w/ 12 levels "Jan","Feb","Mar",..: 4 5 6 7 8 9 10 11 12 1 ...
```

First interruption.

```
# Binary variable to flag the before and after first interruption
df$after_covid_start <- ifelse(df$period == "pre", 0, 1)
# sanity check
table(df$period, df$after_covid_start)
```

```
## 
##              0  1
##   impacted  0 37
##   post      0 24
##   pre      59  0
```

```
# adding a counter column to months since the first interruption (the start of COVID)
Covid_start_index <- min(df$month_index[df$after_covid_start==1])
#Covid_start_index
df$months_since_covid_start <- ifelse(
  df$after_covid_start ==1,
  df$month_index - Covid_start_index + 1,
  0
)
```

Second Interruption (end of COVID)

```
df$after_covid_end <- ifelse(df$period == "post", 1, 0)
table(df$period, df$after_covid_end)
```

```
## 
##              0  1
##   impacted 37  0
##   post      0 24
##   pre      59  0
```

```
covid_end_index <- min(df$month_index[df$after_covid_end == 1])
#covid_end_index

df$months_since_covid_end <- ifelse(

  df$after_covid_end == 1,
  df$month_index - covid_end_index + 1,
  0
)
```