Practical

2025-11-10

Load packages

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.2.0
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(here)
```

```
## here() starts at /home/myke/Desktop/Intro2HDS_R_WEEK5_MAIN
```

Read dataset

```r
hip_data <- read_csv(here("INPUTS/WEEK6/Hip Replacement CCG 1819.csv"))
```

```
## Rows: 28920 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): Provider Code, Procedure, Year, Age Band, Gender
## dbl (76): Revision Flag, Pre-Op Q Assisted, Pre-Op Q Assisted By, Pre-Op Q S...
##
## 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.
```

```r
head(hip_data)
```

```
## # A tibble: 6 x 81
##   `Provider Code` Procedure      `Revision Flag` Year    `Age Band` Gender
##   <chr>           <chr>                    <dbl> <chr>   <chr>      <chr>
## 1 00C             Hip Replacement              0 2018/19 *          *
## 2 00C             Hip Replacement              0 2018/19 *          *
```

```
## 3 00C               Hip Replacement                    1 2018/19 *          *
## 4 00C               Hip Replacement                    1 2018/19 *          *
## 5 00C               Hip Replacement                    0 2018/19 *          *
## 6 00C               Hip Replacement                    0 2018/19 *          *
## # i 75 more variables: 'Pre-Op Q Assisted' <dbl>, 'Pre-Op Q Assisted By' <dbl>,
## #   'Pre-Op Q Symptom Period' <dbl>, 'Pre-Op Q Previous Surgery' <dbl>,
## #   'Pre-Op Q Living Arrangements' <dbl>, 'Pre-Op Q Disability' <dbl>,
## #   'Heart Disease' <dbl>, 'High Bp' <dbl>, Stroke <dbl>, Circulation <dbl>,
## #   'Lung Disease' <dbl>, Diabetes <dbl>, 'Kidney Disease' <dbl>,
## #   'Nervous System' <dbl>, 'Liver Disease' <dbl>, Cancer <dbl>,
## #   Depression <dbl>, Arthritis <dbl>, 'Pre-Op Q Mobility' <dbl>, ...
```

Drop Gender == *

```r
hip_data_gender_valid <- hip_data %>%
  filter(Gender != "*")

head(hip_data_gender_valid)
```

```
## # A tibble: 6 x 81
##   'Provider Code' Procedure      'Revision Flag' Year    'Age Band' Gender
##   <chr>           <chr>                    <dbl> <chr>   <chr>      <chr>
## 1 00C             Hip Replacement              0 2018/19 60 to 69   1
## 2 00C             Hip Replacement              0 2018/19 60 to 69   1
## 3 00C             Hip Replacement              0 2018/19 60 to 69   1
## 4 00C             Hip Replacement              0 2018/19 60 to 69   1
## 5 00C             Hip Replacement              0 2018/19 60 to 69   1
## 6 00C             Hip Replacement              0 2018/19 60 to 69   1
## # i 75 more variables: 'Pre-Op Q Assisted' <dbl>, 'Pre-Op Q Assisted By' <dbl>,
## #   'Pre-Op Q Symptom Period' <dbl>, 'Pre-Op Q Previous Surgery' <dbl>,
## #   'Pre-Op Q Living Arrangements' <dbl>, 'Pre-Op Q Disability' <dbl>,
## #   'Heart Disease' <dbl>, 'High Bp' <dbl>, Stroke <dbl>, Circulation <dbl>,
## #   'Lung Disease' <dbl>, Diabetes <dbl>, 'Kidney Disease' <dbl>,
## #   'Nervous System' <dbl>, 'Liver Disease' <dbl>, Cancer <dbl>,
## #   Depression <dbl>, Arthritis <dbl>, 'Pre-Op Q Mobility' <dbl>, ...
```

Select EQ5D criteria with gender

```r
gender_EQ5D <- hip_data_gender_valid %>%
  select(`Gender`,`Pre-Op Q EQ5D Index`,`Post-Op Q EQ5D Index`) %>%
  rename(Gender  = `Gender`,
        EQ5D_Pre = `Pre-Op Q EQ5D Index`,
        EQ5D_Post = `Post-Op Q EQ5D Index`
        )

head(gender_EQ5D)
```

```
## # A tibble: 6 x 3
##   Gender EQ5D_Pre EQ5D_Post
```

```
##    <chr>        <dbl>        <dbl>
## 1 1           -0.016        0.516
## 2 1            0.159        0.743
## 3 1            0.03         0.727
## 4 1            0.587        0.85
## 5 1            0.623        0.796
## 6 1            0.691        1
```

summary(gender_EQ5D)

```
##      Gender            EQ5D_Pre            EQ5D_Post
##  Length:26611      Min.   :-0.5940     Min.    :-0.5940
##  Class :character  1st Qu.: 0.0300     1st Qu.: 0.6910
##  Mode  :character  Median : 0.3640     Median : 0.8150
##                    Mean   : 0.3374     Mean    : 0.7994
##                    3rd Qu.: 0.6230     3rd Qu.: 1.0000
##                    Max.   : 1.0000     Max.    : 1.0000
##                    NA's   :1676        NA's    :998
```

Clean gender__EQ5D values

```
gender_EQ5D_noNA <- gender_EQ5D %>%
  drop_na()
```

head(gender_EQ5D_noNA)

```
## # A tibble: 6 x 3
##   Gender EQ5D_Pre EQ5D_Post
##   <chr>      <dbl>      <dbl>
## 1 1         -0.016      0.516
## 2 1          0.159      0.743
## 3 1          0.03       0.727
## 4 1          0.587      0.85
## 5 1          0.623      0.796
## 6 1          0.691      1
```

summary(gender_EQ5D_noNA)

```
##      Gender            EQ5D_Pre            EQ5D_Post
##  Length:24042      Min.   :-0.594      Min.    :-0.5940
##  Class :character  1st Qu.: 0.055      1st Qu.: 0.6910
##  Mode  :character  Median : 0.516      Median : 0.8150
##                    Mean   : 0.339      Mean    : 0.7995
##                    3rd Qu.: 0.656      3rd Qu.: 1.0000
##                    Max.   : 1.000      Max.    : 1.0000
```

Convert to Tidy data
```

```
head(gender_EQ5D_noNA)
```

```
## # A tibble: 6 x 3
##   Gender EQ5D_Pre EQ5D_Post
##   <chr>     <dbl>     <dbl>
## 1 1        -0.016     0.516
## 2 1         0.159     0.743
## 3 1         0.03      0.727
## 4 1         0.587     0.85
## 5 1         0.623     0.796
## 6 1         0.691     1
```

```
tidy_gender_EQ5D_noNA <- gender_EQ5D_noNA %>%
  pivot_longer(c(EQ5D_Pre, EQ5D_Post),
               names_to = 'Time',
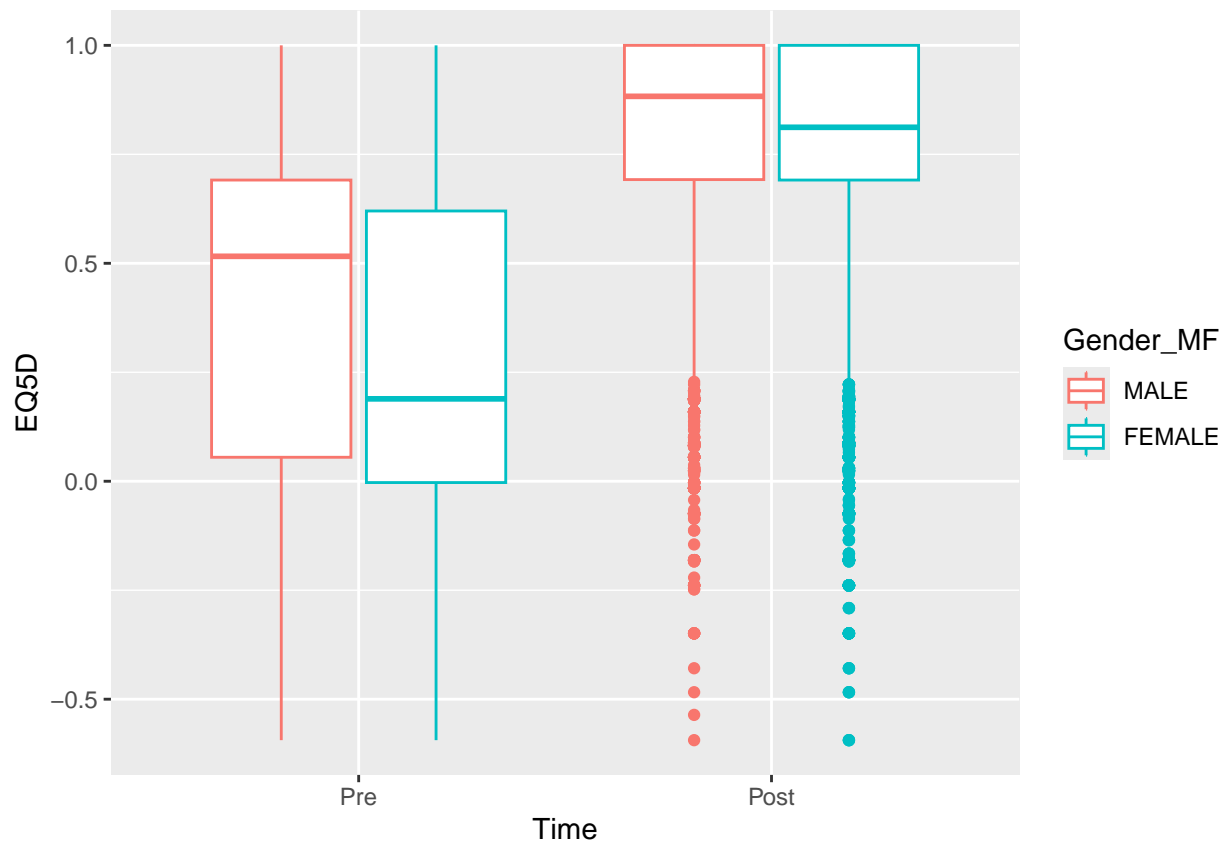               names_prefix = 'EQ5D_',
               values_to = 'EQ5D'
               )

head(tidy_gender_EQ5D_noNA)
```

```
## # A tibble: 6 x 3
##   Gender Time    EQ5D
##   <chr>  <chr>  <dbl>
## 1 1      Pre   -0.016
## 2 1      Post   0.516
## 3 1      Pre    0.159
## 4 1      Post   0.743
## 5 1      Pre    0.03
## 6 1      Post   0.727
```

1. Plot 'EQ-5D Index' scores pre and post operation for each gender

```
tidy_gender_EQ5D_noNA <- mutate(tidy_gender_EQ5D_noNA, Gender_MF = ifelse(Gender == 1, "MALE", "FEMALE")

tidy_gender_EQ5D_noNA$Gender_MF <- factor(tidy_gender_EQ5D_noNA$Gender_MF, levels = c("MALE", "FEMALE")
tidy_gender_EQ5D_noNA$Time <- factor(tidy_gender_EQ5D_noNA$Time, levels = c('Pre', 'Post'))

tidy_gender_EQ5D_noNA %>%
  ggplot() +
  geom_boxplot(aes(x = Time, y = EQ5D, colour = Gender_MF))
```

2. Calculate how many patients in this dataset have been told by a doctor that they have problems caused by a stroke

```
head(hip_data)
```

```
## # A tibble: 6 x 81
##   'Provider Code' Procedure        'Revision Flag' Year    'Age Band' Gender
##   <chr>           <chr>                      <dbl> <chr>   <chr>      <chr>
## 1 00C             Hip Replacement                0 2018/19 *          *
## 2 00C             Hip Replacement                0 2018/19 *          *
## 3 00C             Hip Replacement                1 2018/19 *          *
## 4 00C             Hip Replacement                1 2018/19 *          *
## 5 00C             Hip Replacement                0 2018/19 *          *
## 6 00C             Hip Replacement                0 2018/19 *          *
## # i 75 more variables: 'Pre-Op Q Assisted' <dbl>, 'Pre-Op Q Assisted By' <dbl>,
## #   'Pre-Op Q Symptom Period' <dbl>, 'Pre-Op Q Previous Surgery' <dbl>,
## #   'Pre-Op Q Living Arrangements' <dbl>, 'Pre-Op Q Disability' <dbl>,
## #   'Heart Disease' <dbl>, 'High Bp' <dbl>, Stroke <dbl>, Circulation <dbl>,
## #   'Lung Disease' <dbl>, Diabetes <dbl>, 'Kidney Disease' <dbl>,
## #   'Nervous System' <dbl>, 'Liver Disease' <dbl>, Cancer <dbl>,
## #   Depression <dbl>, Arthritis <dbl>, 'Pre-Op Q Mobility' <dbl>, ...
```

```r
hip_data %>%
  group_by(Stroke) %>%
  summarise(mean(`Pre-Op Q Assisted`), .groups = "drop")
```

```
## # A tibble: 2 x 2
##   Stroke `mean(\`Pre-Op Q Assisted\`)`
##    <dbl>                         <dbl>
## 1      1                          1.74
## 2      9                          1.93
```

```r
hip_data %>%
  filter(Stroke == 1) %>%
  summarise(Count = n())
```

```
## # A tibble: 1 x 1
##   Count
##   <int>
## 1   400
```

3. Create a clean and tidy table with pre and post operation activity levels

```r
head(hip_data)
```

```
## # A tibble: 6 x 81
##   `Provider Code` Procedure       `Revision Flag` Year     `Age Band` Gender
##   <chr>           <chr>                     <dbl> <chr>    <chr>      <chr>
## 1 00C             Hip Replacement               0 2018/19  *          *
## 2 00C             Hip Replacement               0 2018/19  *          *
## 3 00C             Hip Replacement               1 2018/19  *          *
## 4 00C             Hip Replacement               1 2018/19  *          *
## 5 00C             Hip Replacement               0 2018/19  *          *
## 6 00C             Hip Replacement               0 2018/19  *          *
## # i 75 more variables: `Pre-Op Q Assisted` <dbl>, `Pre-Op Q Assisted By` <dbl>,
## #   `Pre-Op Q Symptom Period` <dbl>, `Pre-Op Q Previous Surgery` <dbl>,
## #   `Pre-Op Q Living Arrangements` <dbl>, `Pre-Op Q Disability` <dbl>,
## #   `Heart Disease` <dbl>, `High Bp` <dbl>, Stroke <dbl>, Circulation <dbl>,
## #   `Lung Disease` <dbl>, Diabetes <dbl>, `Kidney Disease` <dbl>,
## #   `Nervous System` <dbl>, `Liver Disease` <dbl>, Cancer <dbl>,
## #   Depression <dbl>, Arthritis <dbl>, `Pre-Op Q Mobility` <dbl>, ...
```

```r
hip_data %>%
  select(contains("activity", ignore.case = TRUE))
```

```
## # A tibble: 28,920 x 2
##    `Pre-Op Q Activity` `Post-Op Q Activity`
##                  <dbl>                <dbl>
## 1                    9                    2
## 2                    3                    9
## 3                    9                    3
```

```
##  4                    3                  1
##  5                    3                  2
##  6                    2                  2
##  7                    2                  1
##  8                    2                  1
##  9                    2                  1
## 10                    2                  2
## # i 28,910 more rows
```

```r
tidy_activity_hip_data <- hip_data %>%
  filter(`Pre-Op Q Activity` != 9, `Post-Op Q Activity` != 9) %>%
  drop_na() %>%
  pivot_longer(c(`Pre-Op Q Activity`, `Post-Op Q Activity`),
               names_to = 'Time',
               values_to = 'Activity',
               values_drop_na = TRUE,
               names_pattern = "(.*)-Op Q Activity",
               ) %>%
  select(Time, Activity) # Only show Time, and Activity for concise overview

head(tidy_activity_hip_data)
```

```
## # A tibble: 6 x 2
##    Time   Activity
##    <chr>     <dbl>
## 1 Pre           3
## 2 Post          1
## 3 Pre           3
## 4 Post          2
## 5 Pre           2
## 6 Post          2
```