Practical Data Cleaning with R

this practice is from PU5058, PU5063 (2024 -25): Introduction to Health Data Science Practical: Data Cleaning with R week 6

there are 3 questions

1. Plot 'EQ-5D Index' scores pre and post operation for each gender
2. Calculate how many patients in this dataset have been told by a doctor that they have problems caused by a stroke
3. Create a clean and tidy table with pre and post operation activity levels

Load packages

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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
```

First question, Plot 'EQ-5D Index' scores pre and post operation for each gender

read in data and see the column names

```
raw_hip_data <- read_csv('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.
```

```
head (raw_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>, ...
```

```
names(raw_hip_data)
```

```
##  [1] "Provider Code"
##  [2] "Procedure"
##  [3] "Revision Flag"
##  [4] "Year"
##  [5] "Age Band"
##  [6] "Gender"
##  [7] "Pre-Op Q Assisted"
##  [8] "Pre-Op Q Assisted By"
##  [9] "Pre-Op Q Symptom Period"
## [10] "Pre-Op Q Previous Surgery"
## [11] "Pre-Op Q Living Arrangements"
## [12] "Pre-Op Q Disability"
## [13] "Heart Disease"
## [14] "High Bp"
## [15] "Stroke"
## [16] "Circulation"
## [17] "Lung Disease"
## [18] "Diabetes"
## [19] "Kidney Disease"
## [20] "Nervous System"
## [21] "Liver Disease"
## [22] "Cancer"
## [23] "Depression"
## [24] "Arthritis"
## [25] "Pre-Op Q Mobility"
## [26] "Pre-Op Q Self-Care"
## [27] "Pre-Op Q Activity"
## [28] "Pre-Op Q Discomfort"
## [29] "Pre-Op Q Anxiety"
## [30] "Pre-Op Q EQ5D Index Profile"
## [31] "Pre-Op Q EQ5D Index"
```

```
## [32] "Post-Op Q Assisted"
## [33] "Post-Op Q Assisted By"
## [34] "Post-Op Q Living Arrangements"
## [35] "Post-Op Q Disability"
## [36] "Post-Op Q Mobility"
## [37] "Post-Op Q Self-Care"
## [38] "Post-Op Q Activity"
## [39] "Post-Op Q Discomfort"
## [40] "Post-Op Q Anxiety"
## [41] "Post-Op Q Satisfaction"
## [42] "Post-Op Q Sucess"
## [43] "Post-Op Q Allergy"
## [44] "Post-Op Q Bleeding"
## [45] "Post-Op Q Wound"
## [46] "Post-Op Q Urine"
## [47] "Post-Op Q Further Surgery"
## [48] "Post-Op Q Readmitted"
## [49] "Post-Op Q EQ5D Index Profile"
## [50] "Post-Op Q EQ5D Index"
## [51] "Hip Replacement EQ5D Index Post-Op Q Predicted"
## [52] "Pre-Op Q EQ VAS"
## [53] "Post-Op Q EQ VAS"
## [54] "Hip Replacement EQ VAS Post-Op Q Predicted"
## [55] "Hip Replacement Pre-Op Q Pain"
## [56] "Hip Replacement Pre-Op Q Sudden Pain"
## [57] "Hip Replacement Pre-Op Q Night Pain"
## [58] "Hip Replacement Pre-Op Q Washing"
## [59] "Hip Replacement Pre-Op Q Transport"
## [60] "Hip Replacement Pre-Op Q Dressing"
## [61] "Hip Replacement Pre-Op Q Shopping"
## [62] "Hip Replacement Pre-Op Q Walking"
## [63] "Hip Replacement Pre-Op Q Limping"
## [64] "Hip Replacement Pre-Op Q Stairs"
## [65] "Hip Replacement Pre-Op Q Standing"
## [66] "Hip Replacement Pre-Op Q Work"
## [67] "Hip Replacement Pre-Op Q Score"
## [68] "Hip Replacement Post-Op Q Pain"
## [69] "Hip Replacement Post-Op Q Sudden Pain"
## [70] "Hip Replacement Post-Op Q Night Pain"
## [71] "Hip Replacement Post-Op Q Washing"
## [72] "Hip Replacement Post-Op Q Transport"
## [73] "Hip Replacement Post-Op Q Dressing"
## [74] "Hip Replacement Post-Op Q Shopping"
## [75] "Hip Replacement Post-Op Q Walking"
## [76] "Hip Replacement Post-Op Q Limping"
## [77] "Hip Replacement Post-Op Q Stairs"
## [78] "Hip Replacement Post-Op Q Standing"
## [79] "Hip Replacement Post-Op Q Work"
## [80] "Hip Replacement Post-Op Q Score"
## [81] "Hip Replacement OHS Post-Op Q Predicted"
```

##remove NA

```
raw_hip_data_noNA <- raw_hip_data %>%
  drop_na() %>%
  filter(Gender!='*') %>%
  rename(EQ5D_Index_PreOp='Pre-Op Q EQ5D Index',EQ5D_Index_PostOp='Post-Op Q EQ5D Index')

head(raw_hip_data_noNA)
```

```
## # 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   2
## # 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 Gender, 'EQ-5D Index' scores pre and post operation. replace 1=male(M), 2= female(F) in Gender column. add patient ID before plotting.

```
Gen_EQ5D_Index <- raw_hip_data_noNA %>%
  select('Gender','EQ5D_Index_PreOp','EQ5D_Index_PostOp') %>%
  mutate(Gender = recode(Gender, `1` = "M", `2` = "F")) %>%
  mutate(patient_ID = row_number())

head(Gen_EQ5D_Index)
```

```
## # A tibble: 6 x 4
##   Gender EQ5D_Index_PreOp EQ5D_Index_PostOp patient_ID
##   <chr>             <dbl>             <dbl>      <int>
## 1 M                -0.016             0.516          1
## 2 M                 0.159             0.743          2
## 3 M                 0.03              0.727          3
## 4 M                 0.587             0.85           4
## 5 M                 0.691             1              5
## 6 F                 0.082             0.848          6
```

```
summary(Gen_EQ5D_Index)
```

```
##     Gender          EQ5D_Index_PreOp  EQ5D_Index_PostOp   patient_ID
##  Length:21718       Min.   :-0.5940   Min.   :-0.5940   Min.   :    1
##  Class :character   1st Qu.: 0.0550   1st Qu.: 0.6910   1st Qu.: 5430
##  Mode  :character   Median : 0.5160   Median : 0.8480   Median :10860
```

```
##                     Mean   : 0.3438   Mean   : 0.8028   Mean   :10860
##                     3rd Qu.: 0.6560   3rd Qu.: 1.0000   3rd Qu.:16289
##                     Max.   : 1.0000   Max.   : 1.0000   Max.   :21718
```

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

```
tidy_Gen_EQ5D_Index <- Gen_EQ5D_Index %>%
  pivot_longer(
    c('EQ5D_Index_PreOp','EQ5D_Index_PostOp'),
    names_to= 'Time',
    names_prefix='EQ5D_Index_',
    values_to= 'EQ5D_Index'
    )

head(tidy_Gen_EQ5D_Index)
```

```
## # A tibble: 6 x 4
##   Gender patient_ID Time   EQ5D_Index
##   <chr>       <int> <chr>       <dbl>
## 1 M               1 PreOp      -0.016
## 2 M               1 PostOp      0.516
## 3 M               2 PreOp       0.159
## 4 M               2 PostOp      0.743
## 5 M               3 PreOp       0.03
## 6 M               3 PostOp      0.727
```

```
summary(tidy_Gen_EQ5D_Index)
```

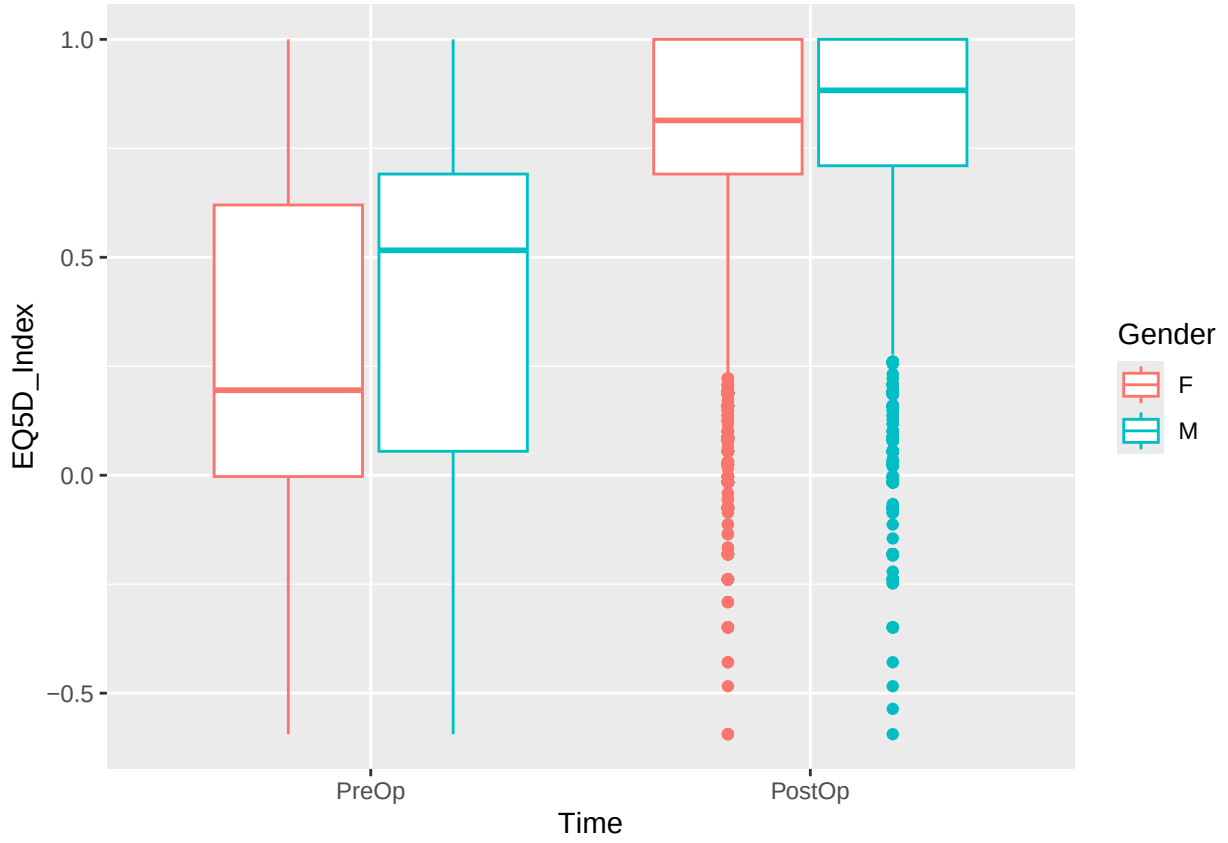
```
##     Gender            patient_ID        Time             EQ5D_Index
##  Length:43436       Min.   :    1   Length:43436       Min.   :-0.5940
##  Class :character   1st Qu.: 5430   Class :character   1st Qu.: 0.1890
##  Mode  :character   Median :10860   Mode  :character   Median : 0.6910
##                     Mean   :10860                      Mean   : 0.5733
##                     3rd Qu.:16289                      3rd Qu.: 0.8480
##                     Max.   :21718                      Max.   : 1.0000
```

change order of Time and visualization of the frame!

```
tidy_Gen_EQ5D_Index$Time <- factor(tidy_Gen_EQ5D_Index$Time, levels=c('PreOp','PostOp'))

tidy_Gen_EQ5D_Index %>%

  ggplot() +
  geom_boxplot(aes(x = Time, y = EQ5D_Index, colour = Gender))
```



Second, calculate how many patients in this dataset have been told by a doctor that they have problems caused by a stroke

```
patient_stroke <- raw_hip_data_noNA %>%
  rename(heart_disease='Heart Disease') %>%
  filter(Stroke==1)

count(patient_stroke)
```

```
## # A tibble: 1 x 1
##       n
##   <int>
## 1   291
```

```
n1 <- nrow(patient_stroke)
```

There are 291 patients in this dataset have been told by a doctor that they have problems caused by a stroke.

Third, create a clean and tidy table with pre and post operation activity levels

```
pre_and_post_act <- raw_hip_data_noNA %>%
  select('Gender','Pre-Op Q Activity', 'Post-Op Q Activity') %>%
  mutate(patient_ID=row_number())
head(pre_and_post_act)
```

```
## # A tibble: 6 x 4
##   Gender `Pre-Op Q Activity` `Post-Op Q Activity` patient_ID
##   <chr>                <dbl>                <dbl>      <int>
## 1 1                        2                    2          1
## 2 1                        2                    2          2
## 3 1                        3                    1          3
## 4 1                        2                    1          4
## 5 1                        2                    1          5
## 6 2                        2                    1          6
```

```
summary(pre_and_post_act)
```

```
##     Gender          Pre-Op Q Activity Post-Op Q Activity   patient_ID
##  Length:21718       Min.   :1.000     Min.   :1.00       Min.   :    1
##  Class :character   1st Qu.:2.000     1st Qu.:1.00       1st Qu.: 5430
##  Mode  :character   Median :2.000     Median :1.00       Median :10860
##                     Mean   :2.132     Mean   :1.43       Mean   :10860
##                     3rd Qu.:2.000     3rd Qu.:2.00       3rd Qu.:16289
##                     Max.   :3.000     Max.   :3.00       Max.   :21718
```

```
tidy_pre_post_act <- pre_and_post_act %>%
  rename(PreOp='Pre-Op Q Activity',PostOp='Post-Op Q Activity') %>%
  pivot_longer(c(PreOp,PostOp),
               names_to='Time',
               values_to='Activity')
head(tidy_pre_post_act)
```

```
## # A tibble: 6 x 4
##   Gender patient_ID Time   Activity
##   <chr>       <int> <chr>     <dbl>
## 1 1               1 PreOp         2
## 2 1               1 PostOp        2
## 3 1               2 PreOp         2
## 4 1               2 PostOp        2
## 5 1               3 PreOp         3
## 6 1               3 PostOp        1
```

```
summary(tidy_pre_post_act)
```

```
##     Gender            patient_ID        Time              Activity
##  Length:43436       Min.   :    1   Length:43436       Min.   :1.000
##  Class :character   1st Qu.: 5430   Class :character   1st Qu.:1.000
##  Mode  :character   Median :10860   Mode  :character   Median :2.000
##                     Mean   :10860                      Mean   :1.781
##                     3rd Qu.:16289                      3rd Qu.:2.000
##                     Max.   :21718                      Max.   :3.000
```