Project 4: Trabeculectomy Outcomes – Data Cleaning

1. Load the data

```
# Read the CSV file
raw_data <- read_csv("data/Trabec pts.csv") %>%
  select(-c(Name, 'Hosp. No....4', 'Hosp. No....21')) %>%
  mutate(Patient_ID = paste0("PX", str_pad(row_number(), 3, pad = "0")))

# Peek at the structure
glimpse(raw_data)
```

```
## Rows: 33
## Columns: 26
## $ ‘Surgery date...1‘    <chr> "2/8/2022", "13/9/2022", "29/11/2022", "13/12/20~
## $ Age                   <dbl> 35, 80, 45, 75, 62, 48, 56, 56, 56, 54, 58, 62, ~
## $ Eye                   <chr> "RE", "RE", "RE", "LE", "RE", "RE", "RE", "LE", ~
## $ Surgery               <chr> "RICCE + Trab.", "RICCE + Trab.", "Trab + Sics",~
## $ Sex                   <chr> "F", "F", "M", "f", "F", "F", "M", "M", "M", "F"~
## $ ‘VA(First Visit)‘     <chr> "folder", "folder", "PL", "1.60", "6.60", "PL", ~
## $ ‘VA(Last visit)...9‘  <chr> NA, NA, "PL", "1.60", "6.60", "6.12", "6.60", "h~
## $ ‘VA(Last visit)...10‘ <chr> NA, NA, "PL", "1.60", "6.60", "6.12", "6.60", "h~
## $ ‘IOP(Last Visit)‘     <chr> NA, NA, "16.00", "9.00", "32.00", "10.00", "14.0~
## $ ‘Current medication‘  <chr> NA, NA, NA, "hypromellose", "brimonidine and dor~
## $ ‘VA(1DPO)‘            <chr> NA, NA, "PL", "hm", "6.36", "6.60", NA, NA, "6.1~
## $ ‘IOP(1DPO)‘           <chr> NA, NA, "n/a", "2.00", "12.00", "n/a", NA, NA, "~
## $ ...15                 <chr> NA, NA, "phacomorphic glaucoma", NA, NA, "???", ~
## $ ‘Pre-Op VA‘           <chr> NA, NA, "PL", "hm", "6.60", "PL", NA, NA, "6.12"~
## $ ‘Pre-Op IOP‘          <chr> NA, NA, "30", "11", "32", "n/a", NA, NA, "12", "~
## $ ‘Pre-Op medication‘   <chr> NA, NA, "IV Mannitol, Diamox", "misopt", "misopt~
## $ ‘First visit date‘    <chr> NA, NA, "28/11/2022", "9/28/2016", "30/09/2021",~
## $ ‘last visit date‘     <chr> NA, NA, NA, "2/7/2024", "30/07/2025", NA, NA, "1~
## $ ‘Surgery date...22‘   <chr> "2/8/2022", "13/9/2022", "29/11/2022", "13/12/20~
## $ ‘6/12 post op IOP‘    <chr> "n/a", "n/a", "n/a", "6", "55", "n/a", "10", "11~
## $ ‘6/12 post op VA‘     <chr> "n/a", "n/a", "n/a", "1.60", "6.36", "n/a", "5.6~
## $ ‘1 yr post op IOP‘    <chr> "n/a", "n/a", "n/a", "9", "33", "n/a", "n/a", "n~
## $ ‘1 yr post op VA‘     <chr> "n/a", "n/a", "n/a", "1.60", "4.60", "n/a", "n/a~
## $ ‘2 yr post op IOP‘    <chr> "n/a", "n/a", "n/a", "n/a", "n/a", "n/a", "n/a",~
## $ ‘2 yr post op VA‘     <chr> "n/a", "n/a", "n/a", "n/a", "n/a", "n/a", "n/a",~
## $ Patient_ID            <chr> "PX001", "PX002", "PX003", "PX004", "PX005", "PX~
```

2. Rename relevant columns

```r
data <- raw_data %>%
  mutate(
    va_preop = `VA(First Visit)`,
    va_1dpo = `VA(1DPO)`,
    va_6mo = `6/12 post op VA`,
    va_1yr = `1 yr post op VA`,
    iop_preop = `Pre-Op IOP`,
    iop_1dpo= `IOP(1DPO)`,
    iop_6mo = `6/12 post op IOP`,
    iop_1yr = `1 yr post op IOP`
  )
head(data)
```

```
## # A tibble: 6 x 34
##   `Surgery date...1`   Age Eye   Surgery       Sex   `VA(First Visit)`
##   <chr>              <dbl> <chr> <chr>         <chr> <chr>
## 1 2/8/2022              35 RE    RICCE + Trab. F     folder
## 2 13/9/2022             80 RE    RICCE + Trab. F     folder
## 3 29/11/2022            45 RE    Trab + Sics   M     PL
## 4 13/12/2022            75 LE    trabs sics    f     1.60
## 5 17/1/2023             62 RE    Trab + Sics   F     6.60
## 6 17/1/2023             48 RE    Trab + Sics   F     PL
## # i 28 more variables: `VA(Last visit)...9` <chr>, `VA(Last visit)...10` <chr>,
## #   `IOP(Last Visit)` <chr>, `Current medication` <chr>, `VA(1DPO)` <chr>,
## #   `IOP(1DPO)` <chr>, ...15 <chr>, `Pre-Op VA` <chr>, `Pre-Op IOP` <chr>,
## #   `Pre-Op medication` <chr>, `First visit date` <chr>,
## #   `last visit date` <chr>, `Surgery date...22` <chr>,
## #   `6/12 post op IOP` <chr>, `6/12 post op VA` <chr>,
## #   `1 yr post op IOP` <chr>, `1 yr post op VA` <chr>, ...
```

3. Normalize date

3.1 normalize_dates() Function

```r
library(lubridate)

normalize_dates <- function(x) {
  x_clean <- parse_date_time(x,
                             orders = c("dmy", "mdy", "ymd", "dmy HMS", "ymd HMS"),
                             tz = "UTC",
                             exact = FALSE)
  return(as.Date(x_clean))
}
```

3.2 Normalize

```r
data <- data %>%
  mutate(
    surgery_date = normalize_dates(`Surgery date...1`),
```

```
    last_visit_date = normalize_dates(`last visit date`),
    first_visit_date = normalize_dates(`First visit date`)
  )
head(data)
```

```
## # A tibble: 6 x 37
##   `Surgery date...1`   Age Eye   Surgery       Sex   `VA(First Visit)`
##   <chr>              <dbl> <chr> <chr>         <chr> <chr>
## 1 2/8/2022              35 RE    RICCE + Trab. F     folder
## 2 13/9/2022             80 RE    RICCE + Trab. F     folder
## 3 29/11/2022            45 RE    Trab + Sics   M     PL
## 4 13/12/2022            75 LE    trabs sics    f     1.60
## 5 17/1/2023             62 RE    Trab + Sics   F     6.60
## 6 17/1/2023             48 RE    Trab + Sics   F     PL
## # i 31 more variables: `VA(Last visit)...9` <chr>, `VA(Last visit)...10` <chr>,
## #   `IOP(Last Visit)` <chr>, `Current medication` <chr>, `VA(1DPO)` <chr>,
## #   `IOP(1DPO)` <chr>, ...15 <chr>, `Pre-Op VA` <chr>, `Pre-Op IOP` <chr>,
## #   `Pre-Op medication` <chr>, `First visit date` <chr>,
## #   `last visit date` <chr>, `Surgery date...22` <chr>,
## #   `6/12 post op IOP` <chr>, `6/12 post op VA` <chr>,
## #   `1 yr post op IOP` <chr>, `1 yr post op VA` <chr>, ...
```

3.3 Normalize Surgery Column

```
# Clean and standardize surgery types
data <- data %>%
  mutate(
    Surgery = str_to_lower(trimws(Surgery)),
    surgery_clean = case_when(
      str_detect(Surgery, "ricce") ~ "RICCE + Trab",
      str_detect(Surgery, "sics") & str_detect(Surgery, "trab") ~ "Trab + SICS",
      str_detect(Surgery, "sics") ~ "SICS",
      str_detect(Surgery, "trab") ~ "Trabeculectomy",
      TRUE ~ NA_character_
    ),
    surgery_clean = factor(surgery_clean)
  )
```

4. Cleaning up va

4.1 va_to_logmar function

```
va_to_logmar <- function(x) {
  x_clean <- tolower(trimws(as.character(x)))

  # 1. Map non-Snellen textual VA to approximate logMAR
  replacements <- c(
    "pl"             = 2.0,
    "hm"             = 2.3,
```

```
    "cf"             = 1.9,
    "npl"            = 2.7,
    "nlp"            = 2.7,
    "no perception"  = 2.7,
    "no lp"          = 2.7,
    "nil"            = NA,
    "folder"         = NA
  )

  x_clean <- ifelse(x_clean %in% names(replacements),
                    replacements[x_clean],
                    x_clean)

  # 2. Convert dot-format to Snellen (e.g. 6.36 -> 6/36, 3.60 -> 3/60)
  x_clean <- gsub("^(\\d+)\\.(\\d+)$", "\\1/\\2", x_clean)

  # 3. Convert to logMAR
  logmar <- suppressWarnings(vapply(x_clean, function(val) {
    # 3a. If already numeric and plausible, treat as logMAR
    num_val <- suppressWarnings(as.numeric(val))
    if (!is.na(num_val)) {
      # accept only plausible logMAR range
      if (num_val >= 0 && num_val <= 3) {
        return(num_val)
      } else {
        return(NA_real_)
      }
    }

    # 3b. If Snellen "a/b", use a and b
    if (grepl("^[0-9]+/[0-9]+$", val)) {
      parts  <- strsplit(val, "/")[[1]]
      num    <- as.numeric(parts[1])
      denom  <- as.numeric(parts[2])

      if (is.na(num) || is.na(denom) || num <= 0) {
        return(NA_real_)
      }

      # logMAR = log10(denominator / numerator)
      return(round(log10(denom / num), 2))
    }

    # 3c. Everything else → NA
    return(NA_real_)
  }, numeric(1)))

  as.numeric(logmar)
}
```

4.2 code

```
data <- data %>%
  mutate(
    va_preop_logmar = va_to_logmar(va_preop),
    va_1dpo_logmar = va_to_logmar(va_1dpo),
    va_6mo_logmar = va_to_logmar(va_6mo),
    va_1yr_logmar = va_to_logmar(va_1yr)
  )
head(data)
```

```
## # A tibble: 6 x 42
##   `Surgery date...1`   Age Eye   Surgery       Sex   `VA(First Visit)`
##   <chr>              <dbl> <chr> <chr>         <chr> <chr>
## 1 2/8/2022              35 RE    ricce + trab. F     folder
## 2 13/9/2022             80 RE    ricce + trab. F     folder
## 3 29/11/2022            45 RE    trab + sics   M     PL
## 4 13/12/2022            75 LE    trabs sics    f     1.60
## 5 17/1/2023             62 RE    trab + sics   F     6.60
## 6 17/1/2023             48 RE    trab + sics   F     PL
## # i 36 more variables: `VA(Last visit)...9` <chr>, `VA(Last visit)...10` <chr>,
## #   `IOP(Last Visit)` <chr>, `Current medication` <chr>, `VA(1DPO)` <chr>,
## #   `IOP(1DPO)` <chr>, ...15 <chr>, `Pre-Op VA` <chr>, `Pre-Op IOP` <chr>,
## #   `Pre-Op medication` <chr>, `First visit date` <chr>,
## #   `last visit date` <chr>, `Surgery date...22` <chr>,
## #   `6/12 post op IOP` <chr>, `6/12 post op VA` <chr>,
## #   `1 yr post op IOP` <chr>, `1 yr post op VA` <chr>, ...
```

5. Code to Clean IOP Columns

```
# Standardize and convert IOP columns to numeric
clean_iop <- function(x) {
  x <- tolower(trimws(as.character(x)))
  x[x %in% c("na", "nil", "", "-", "--")] <- NA
  as.numeric(x)
}

data <- data %>%
  mutate(
    iop_1dpo  = clean_iop(iop_1dpo),
    iop_6mo  = clean_iop(iop_6mo),
    iop_1yr  = clean_iop(iop_1yr)
  )
```

5.1

```
data <- data %>%
  mutate(
    iop_drop_6mo = iop_1dpo - iop_6mo,
```

```
    iop_drop_1yr = iop_6mo - iop_1yr
  )
head(data)
```

```
## # A tibble: 6 x 44
##   `Surgery date...1`   Age Eye   Surgery       Sex   `VA(First Visit)`
##   <chr>              <dbl> <chr> <chr>         <chr> <chr>
## 1 2/8/2022              35 RE    ricce + trab. F     folder
## 2 13/9/2022             80 RE    ricce + trab. F     folder
## 3 29/11/2022            45 RE    trab + sics   M     PL
## 4 13/12/2022            75 LE    trabs sics    f     1.60
## 5 17/1/2023             62 RE    trab + sics   F     6.60
## 6 17/1/2023             48 RE    trab + sics   F     PL
## # i 38 more variables: `VA(Last visit)...9` <chr>, `VA(Last visit)...10` <chr>,
## #   `IOP(Last Visit)` <chr>, `Current medication` <chr>, `VA(1DPO)` <chr>,
## #   `IOP(1DPO)` <chr>, ...15 <chr>, `Pre-Op VA` <chr>, `Pre-Op IOP` <chr>,
## #   `Pre-Op medication` <chr>, `First visit date` <chr>,
## #   `last visit date` <chr>, `Surgery date...22` <chr>,
## #   `6/12 post op IOP` <chr>, `6/12 post op VA` <chr>,
## #   `1 yr post op IOP` <chr>, `1 yr post op VA` <chr>, ...
```

6. Pivot longer

```
clean_iop <- function(x) {
  x <- tolower(trimws(as.character(x)))
  x[x %in% c("na", "n/a", "nil", "", "-", "--", "n.a.")] <- NA
  suppressWarnings(as.numeric(x))
}

data <- data %>%
  mutate(
    across(
      starts_with("iop_"),
      clean_iop
    )
  )
```

```
# 1) IOP long format
iop_long <- data %>%
  select(Patient_ID, Eye, Age, Sex,
         iop_preop, iop_1dpo, iop_6mo, iop_1yr) %>%
  pivot_longer(
    cols = starts_with("iop_"),
    names_to = "timepoint",
    values_to = "iop"
  ) %>%
  mutate(
    timepoint = factor(
      timepoint,
      levels = c("iop_preop", "iop_1dpo", "iop_6mo", "iop_1yr")
    )
```

```
  )

# 2) VA long format (adjust names to what you actually have)
va_long <- data %>%
  select(Patient_ID, Eye,
         va_preop_logmar, va_1dpo_logmar, va_6mo_logmar, va_1yr_logmar) %>%
  pivot_longer(
    cols = starts_with("va_"),
    names_to = "timepoint",
    values_to = "va_logmar"
  ) %>%
  mutate(
    timepoint = factor(
      timepoint,
      levels = c("va_preop_logmar", "va_1dpo_logmar", "va_6mo_logmar",
                 "va_1yr_logmar")
    )
  )
```

Final Column Cleanup: Drop raw & duplicate fields

```
final_cols <- c(
  "Patient_ID", "Sex", "Age", "Eye", "surgery_clean",
  "surgery_date", "first_visit_date", "last_visit_date", "Pre-Op medication", "Current medication",

  # Cleaned VA and IOP at each timepoint
  "va_preop_logmar", "va_1dpo_logmar", "va_6mo_logmar", "va_1yr_logmar",
  "iop_preop", "iop_1dpo", "iop_6mo", "iop_1yr"
  )

# Retain only final analysis variables
data <- data[, final_cols]
```

7. Save cleaned dataset for analysis

```
# Create a data/ folder if it doesn't exist
if (!dir.exists("data")) dir.create("data")

# Save the fully cleaned trabeculectomy dataset
readr::write_csv(data, "data/trab_clean.csv")
```

```
names(data)
```

```
##  [1] "Patient_ID"         "Sex"                "Age"
##  [4] "Eye"                "surgery_clean"      "surgery_date"
##  [7] "first_visit_date"   "last_visit_date"    "Pre-Op medication"
## [10] "Current medication" "va_preop_logmar"    "va_1dpo_logmar"
## [13] "va_6mo_logmar"      "va_1yr_logmar"      "iop_preop"
## [16] "iop_1dpo"           "iop_6mo"            "iop_1yr"
```