Data Preparation

Berge Sievers

2025-10-09

Part I

For the missing data, the variables does the patient has amnesia for the event and a headache at the time of ED evaluation, which were represented by 91 to indicate a pre-verbal patient or children who haven't started speaking yet. Moreover, when the patient is intubated or otherwise unable to give an understandable verbal response, a non-verbal is marked as 91. On the other hand, 92 in the duration of loss of consciousness and the duration of post-traumatic seizure are not applicable and are to be treated as missing values. This means that for every variable that has a 91 or 92, the values will be replaced with NAs. A variable like the duration of post-traumatic seizure has over 95% of values as 92, which is not applicable; therefore, this variable is removed from the dataset. Additionally, all missing values were removed from the Dataset.

We changed the variable names to make sure that they are understandable and match the information given in the data dictionary.

```
# Load necessary libraries:
# ...
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(stringr)
library(readr)
citbi <- read_csv("C:/Users/ThinkPad/Downloads/citbi.csv")
```

```
## Rows: 30379 Columns: 26
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (26): PatNum, Amnesia_verb, LocLen, Seiz, SeizLen, ActNorm, HA_verb, Vom...
##
## 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
# We preview the dataset
glimpse(citbi)
```

```
## Rows: 30,379
## Columns: 26
## $ PatNum       <dbl> 2986, 29925, 29710, 37529, 2757, 38938, 9642, 31313, 1418~
## $ Amnesia_verb <dbl> 91, 91, 0, 1, 0, NA, 0, 0, 91, 91, 0, 0, 91, 0, 0, 0, 0, ~
## $ LocLen       <dbl> 92, 92, NA, NA, 1, NA, 92, 92, 92, 92, 1, 3, 92, 92, 92, ~
## $ Seiz         <dbl> 0, 0, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ SeizLen      <dbl> 92, 92, 92, 92, 92, 92, 92, 92, 92, 92, 92, 92, 92, 92, 92, 9~
## $ ActNorm      <dbl> 1, 1, NA, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, NA, 1, 1, 1, 0~
## $ HA_verb      <dbl> 91, 91, 1, 0, 1, 0, 0, 0, 0, 91, 0, 0, 91, 0, 1, 1, 0, 1,~
## $ Vomit        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, ~
## $ Dizzy        <dbl> NA, NA, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, NA, 0, 0, 0, 0, 0, ~
## $ GCSEye       <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, ~
## $ GCSVerbal    <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, ~
## $ GCSMotor     <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, ~
## $ GCSTotal     <dbl> 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 15, 1~
## $ AMS          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ SFxPalp      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, ~
## $ FontBulg     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ Hema         <dbl> 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, ~
## $ Clav         <dbl> 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, ~
## $ NeuroD       <dbl> 0, 0, NA, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ OSI          <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, ~
## $ CTForm1      <dbl> 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, ~
## $ AgeInMonth   <dbl> 5, 21, 205, 157, 199, 105, 106, 124, 22, 17, 159, 195, 9,~
## $ Gender       <dbl> 1, 2, 1, 1, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 1, 1, 1, 1, 1, ~
## $ CTDone       <dbl> 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, ~
## $ DeathTBI     <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ PosIntFinal  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

Part I: Data Cleaning

```r
vars_with_91 <- c("Amnesia_verb", "HA_verb")
vars_with_92 <- c("LocLen", "SeizLen")
citbi_clean <- citbi %>%
  mutate(
    across(all_of(vars_with_91), ~ na_if(., 91)),
    across(all_of(vars_with_92), ~ na_if(., 92))
  )
citbi_clean
```

```
## # A tibble: 30,379 x 26
##    PatNum Amnesia_verb LocLen  Seiz SeizLen ActNorm HA_verb Vomit Dizzy GCSEye
##     <dbl>        <dbl>  <dbl> <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl>  <dbl>
##  1   2986           NA     NA     0      NA       1      NA     0    NA      4
##  2  29925           NA     NA     0      NA       1      NA     0    NA      4
##  3  29710            0     NA    NA      NA      NA       1     0     1      4
##  4  37529            1     NA     0      NA       1       0     0     0      4
##  5   2757            0      1     0      NA       1       1     0     0      4
```

```
##  6  38938           NA     NA     0      NA       1       0     0     0      4
##  7   9642            0     NA     0      NA       1       0     0     0      4
##  8  31313            0     NA     0      NA       1       0     0     0      4
##  9  14183           NA     NA     0      NA       0       0     0     0      4
## 10  15180           NA     NA     0      NA       1      NA     0     0      4
## # i 30,369 more rows
## # i 16 more variables: GCSVerbal <dbl>, GCSMotor <dbl>, GCSTotal <dbl>,
## #   AMS <dbl>, SFxPalp <dbl>, FontBulg <dbl>, Hema <dbl>, Clav <dbl>,
## #   NeuroD <dbl>, OSI <dbl>, CTForm1 <dbl>, AgeInMonth <dbl>, Gender <dbl>,
## #   CTDone <dbl>, DeathTBI <dbl>, PosIntFinal <dbl>
```

Renaming variable labels

```
citbi_clean <- citbi_clean %>%
  rename(
    patient_number = PatNum,
    amnesia_event = Amnesia_verb,
    loc_duration = LocLen,
    seizure = Seiz,
    seizure_duration = SeizLen,
    normal_activity = ActNorm,
    headache = HA_verb,
    vomiting = Vomit,
    dizziness = Dizzy,
    gcs_eye = GCSEye,
    gcs_verbal = GCSVerbal,
    gcs_motor = GCSMotor,
    gcs_total = GCSTotal,
    altered_mental_status = AMS,
    skull_fracture_palpable = SFxPalp,
    fontanelle_bulging = FontBulg,
    hematoma = Hema,
    clavicle_trauma = Clav,
    neurological_deficit = NeuroD,
    other_significant_injury = OSI,
    ct_planned = CTForm1,
    age_months = AgeInMonth,
    gender = Gender,
    ct_done = CTDone,
    death_tbi = DeathTBI,
    citbi_outcome = PosIntFinal
  )
```

```
# identification of variable types
## we started by converting the labels to factors
citbi_clean <- citbi_clean %>%
  mutate(
    amnesia_event = factor(amnesia_event,
                           levels = c(0, 1),
                           labels = c("No", "Yes")),
    loc_duration = factor(loc_duration,
```

```
                        levels = c(1, 2, 3, 4),
                        labels = c("<5 sec", "5 sec-<1 min", "1-5 min", ">5 min"),
                        ordered = TRUE),
  seizure = factor(seizure,
                   levels = c(0, 1),
                   labels = c("No", "Yes")),
  seizure_duration = factor(seizure_duration,
                            levels = c(1, 2, 3, 4),
                            labels = c("<1 min", "1-<5 min", "5-15 min", ">15 min"),
                            ordered = TRUE),
  normal_activity = factor(normal_activity,
                           levels = c(0, 1),
                           labels = c("No", "Yes")),
  headache = factor(headache,
                    levels = c(0, 1),
                    labels = c("No", "Yes")),
  vomiting = factor(vomiting,
                    levels = c(0, 1),
                    labels = c("No", "Yes")),
  dizziness = factor(dizziness,
                     levels = c(0, 1),
                     labels = c("No", "Yes")),
  gcs_eye = factor(gcs_eye,
                   levels = c(1, 2, 3, 4),
                   labels = c("None", "Pain", "Verbal", "Spontaneous"),
                   ordered = TRUE),
  gcs_verbal = factor(gcs_verbal,
                      levels = c(1, 2, 3, 4, 5),
                      labels = c("None", "Incomprehensible sounds", "Inappropriate words",
                                 "Confused", "Oriented"),
                      ordered = TRUE),
  gcs_motor = factor(gcs_motor,
                     levels = c(1, 2, 3, 4, 5, 6),
                     labels = c("None", "Abnormal extension", "Abnormal flexure",
                                "Withdraws to pain", "Localizes pain", "Follows commands"),
                     ordered = TRUE),
  altered_mental_status = factor(altered_mental_status,
                                 levels = c(0, 1),
                                 labels = c("No", "Yes")),
  skull_fracture_palpable = factor(skull_fracture_palpable,
                                   levels = c(0, 1, 2),
                                   labels = c("No", "Yes", "Unclear")),
  fontanelle_bulging = factor(fontanelle_bulging,
                              levels = c(0, 1),
                              labels = c("No/Closed", "Yes")),
  hematoma = factor(hematoma,
                    levels = c(0, 1),
                    labels = c("No", "Yes")),
  clavicle_trauma = factor(clavicle_trauma,
                           levels = c(0, 1),
                           labels = c("No", "Yes")),
  neurological_deficit = factor(neurological_deficit,
                                levels = c(0, 1),
```

```
                                    labels = c("No", "Yes")),
    other_significant_injury = factor(other_significant_injury,
                                      levels = c(0, 1),
                                      labels = c("No", "Yes")),
    ct_planned = factor(ct_planned,
                        levels = c(0, 1),
                        labels = c("No", "Yes")),
    gender = factor(gender,
                    levels = c(1, 2),
                    labels = c("Male", "Female")),
    ct_done = factor(ct_done,
                     levels = c(0, 1),
                     labels = c("No", "Yes")),
    death_tbi = factor(death_tbi,
                       levels = c(0, 1),
                       labels = c("No", "Yes")),
    citbi_outcome = factor(citbi_outcome,
                           levels = c(0, 1),
                           labels = c("No", "Yes"))
  )
```

```
# We made sure continuous variables remained as numeric
citbi_clean <- citbi_clean %>%
  mutate(
    patient_number = as.numeric(patient_number),
    age_months = as.numeric(age_months),
    gcs_total = as.numeric(gcs_total)
  )
```

Part II: Exploratory Data Analysis (EDA)

```
missing_summary <- citbi_clean %>%
  summarise(across(everything(),
                   list(n_missing = ~sum(is.na(.)),
                        pct_missing = ~mean(is.na(.)) * 100))) %>%
  pivot_longer(
    cols = everything(),
    names_to = c("variable", ".value"),
    names_sep = "_"
  )
```

```
## Warning: Expected 2 pieces. Additional pieces discarded in 52 rows [1, 2, 3, 4, 5, 6, 7,
## 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, ...].
```

```
missing_summary
```

```
## # A tibble: 37 x 22
##    variable number   event duration     n   pct activity   eye verbal motor
##    <chr>     <dbl>   <dbl>    <dbl> <int> <dbl>    <dbl> <dbl>  <dbl> <dbl>
##  1 patient       0    NA       NA      NA NA       NA       NA     NA    NA
```

```
##  2 patient       0    NA       NA      NA NA       NA       NA     NA    NA
##  3 amnesia      NA 11933       NA      NA NA       NA       NA     NA    NA
##  4 amnesia      NA    39.3     NA      NA NA       NA       NA     NA    NA
##  5 loc          NA    NA    27398      NA NA       NA       NA     NA    NA
##  6 loc          NA    NA       90.2    NA NA       NA       NA     NA    NA
##  7 seizure      NA    NA    30046     630  2.07    NA       NA     NA    NA
##  8 seizure      NA    NA       98.9    NA NA       NA       NA     NA    NA
##  9 normal       NA    NA       NA      NA NA     2324       NA     NA    NA
## 10 normal       NA    NA       NA      NA NA        7.65    NA     NA    NA
## # i 27 more rows
## # i 12 more variables: total <dbl>, mental <dbl>, fracture <dbl>,
## #   bulging <dbl>, trauma <dbl>, deficit <dbl>, significant <dbl>,
## #   planned <dbl>, months <dbl>, done <dbl>, tbi <dbl>, outcome <dbl>
```

```r
# for Numerical Values
# Summary stats for numeric variables
citbi_clean %>%
  select(age_months, gcs_total) %>%
  summary()
```

```
##    age_months       gcs_total
##  Min.   :  0.00   Min.   : 3.00
##  1st Qu.: 24.00   1st Qu.:15.00
##  Median : 68.00   Median :15.00
##  Mean   : 84.46   Mean   :14.84
##  3rd Qu.:144.00   3rd Qu.:15.00
##  Max.   :215.00   Max.   :15.00
```

```r
# for categorical variables
# Select categorical variables
categorical_vars <- c("amnesia_event", "seizure", "normal_activity",
                      "headache", "vomiting", "dizziness", "altered_mental_status",
                      "skull_fracture_palpable", "fontanelle_bulging", "hematoma",
                      "clavicle_trauma", "neurological_deficit", "other_significant_injury",
                      "ct_planned", "ct_done", "death_tbi", "citbi_outcome", "gender")

# Frequency tables
lapply(citbi_clean[categorical_vars], table, useNA = "ifany")
```

```
## $amnesia_event
##
##    No   Yes  <NA>
## 15279  3167 11933
##
## $seizure
##
##    No   Yes  <NA>
## 29342   407   630
##
## $normal_activity
##
##    No   Yes  <NA>
##  4802 23253  2324
```

```
##
## $headache
##
##    No   Yes  <NA>
## 11121  9006 10252
##
## $vomiting
##
##    No   Yes  <NA>
## 25959  4106   314
##
## $dizziness
##
##    No   Yes  <NA>
## 17428  1808 11143
##
## $altered_mental_status
##
##    No   Yes  <NA>
## 25575  4570   234
##
## $skull_fracture_palpable
##
##      No     Yes Unclear    <NA>
##   29469     152     685      73
##
## $fontanelle_bulging
##
## No/Closed       Yes      <NA>
##     30239        25       115
##
## $hematoma
##
##    No   Yes  <NA>
## 18248 11912   219
##
## $clavicle_trauma
##
##    No   Yes  <NA>
## 10810 19465   104
##
## $neurological_deficit
##
##    No   Yes  <NA>
## 29439   453   487
##
## $other_significant_injury
##
##    No   Yes  <NA>
## 27017  3235   127
##
## $ct_planned
##
##    No   Yes  <NA>
```

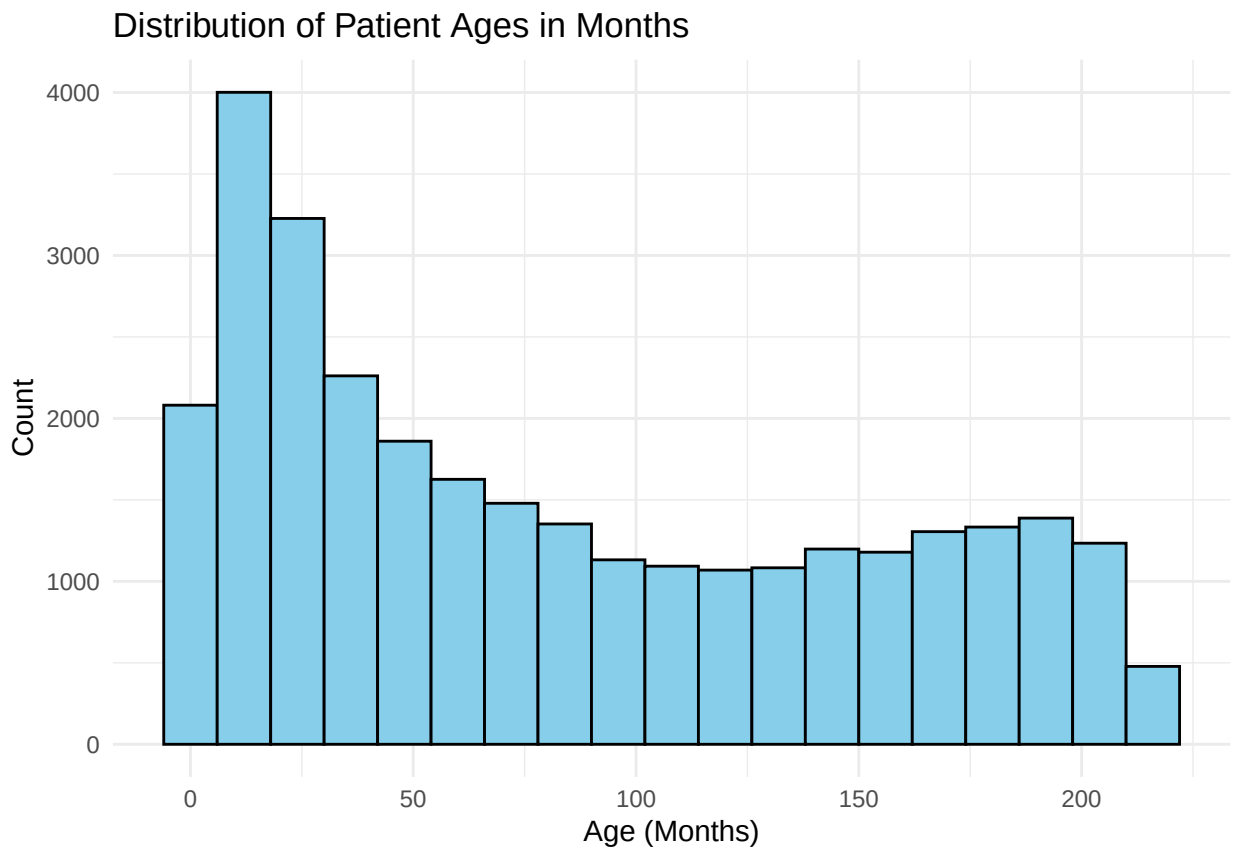
```
## 18549 11811    19
##
## $ct_done
##
##    No   Yes
## 19278 11101
##
## $death_tbi
##
##    No   Yes  <NA>
## 30325    50     4
##
## $citbi_outcome
##
##    No   Yes  <NA>
## 29819   547    13
##
## $gender
##
##   Male Female   <NA>
##  19003  11373      3
```

5

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
# We perform the histogram of patient ages in months
ggplot(citbi_clean, aes(x = age_months)) +
  geom_histogram(binwidth = 12, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Patient Ages in Months",
       x = "Age (Months)", y = "Count") +
  theme_minimal()
```



6

```
summary_table <- citbi_clean %>%
  group_by(loc_duration, citbi_outcome) %>%
  summarise(
    mean_age_months = mean(age_months, na.rm = TRUE),
    mean_gcs_total = mean(gcs_total, na.rm = TRUE),
    count = n()
  ) %>%
  arrange(loc_duration, citbi_outcome)
```

```
## `summarise()` has grouped output by 'loc_duration'. You can override using the
## `.groups` argument.
```

```
summary_table
```

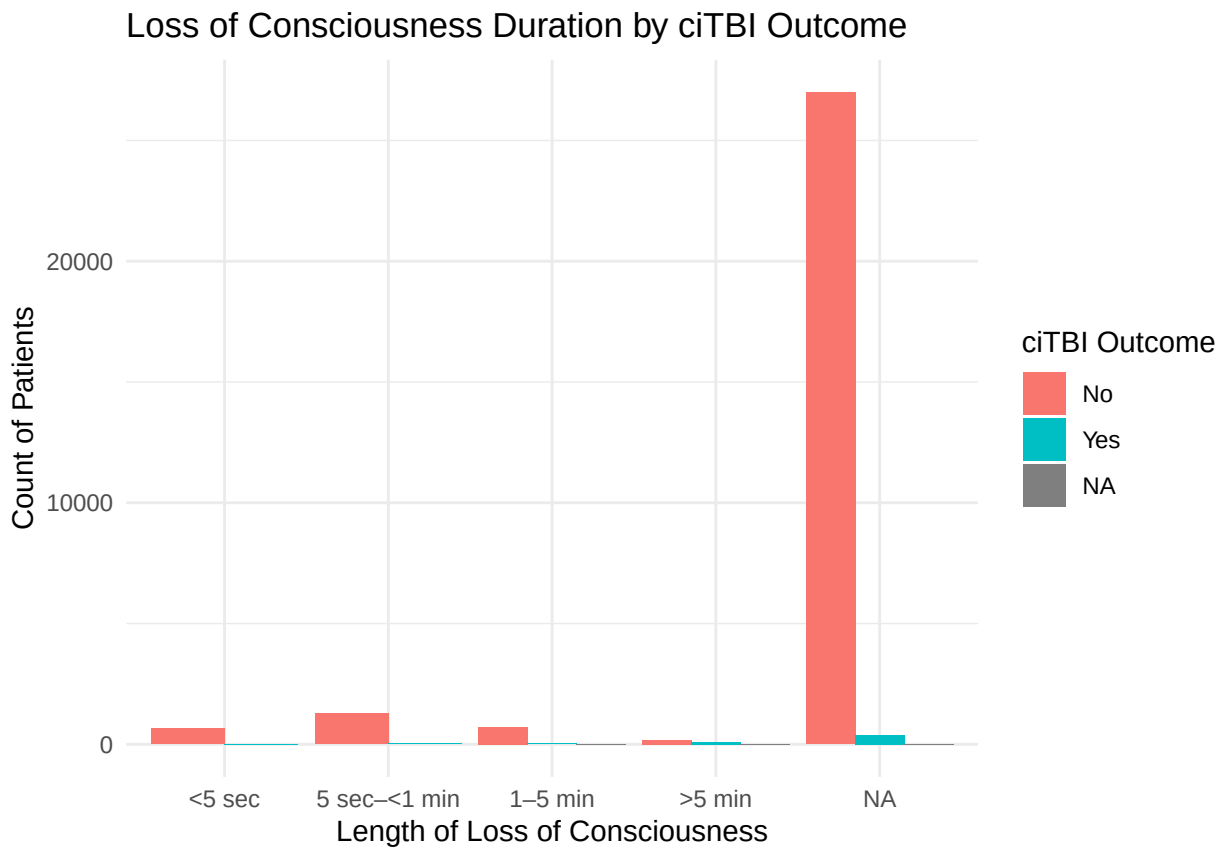
```
## # A tibble: 13 x 5
## # Groups:   loc_duration [5]
##    loc_duration citbi_outcome mean_age_months mean_gcs_total count
##    <ord>        <fct>                   <dbl>          <dbl> <int>
##  1 <5 sec       No                      130.           14.9    652
##  2 <5 sec       Yes                      95.4          14.8      8
```

```
##  3 5 sec-<1 min No                    122.          14.9   1283
##  4 5 sec-<1 min Yes                   117.          14.0     25
##  5 1-5 min      No                    128.          14.8    723
##  6 1-5 min      Yes                   120.          12.7     30
##  7 1-5 min      <NA>                  214           15        1
##  8 >5 min       No                     96.0         13.8    168
##  9 >5 min       Yes                   106.           5.44    90
## 10 >5 min       <NA>                   99            9        1
## 11 <NA>         No                     80.1         14.9  26993
## 12 <NA>         Yes                    93.0         11.7    394
## 13 <NA>         <NA>                  112.          14.5     11
```

7

```r
# Side-by-side bar chart
library(ggplot2)

ggplot(citbi_clean, aes(x = loc_duration, fill = citbi_outcome)) +
  geom_bar(position = "dodge") +
  labs(
    title = "Loss of Consciousness Duration by ciTBI Outcome",
    x = "Length of Loss of Consciousness",
    y = "Count of Patients",
    fill = "ciTBI Outcome"
  ) +
  theme_minimal()
```

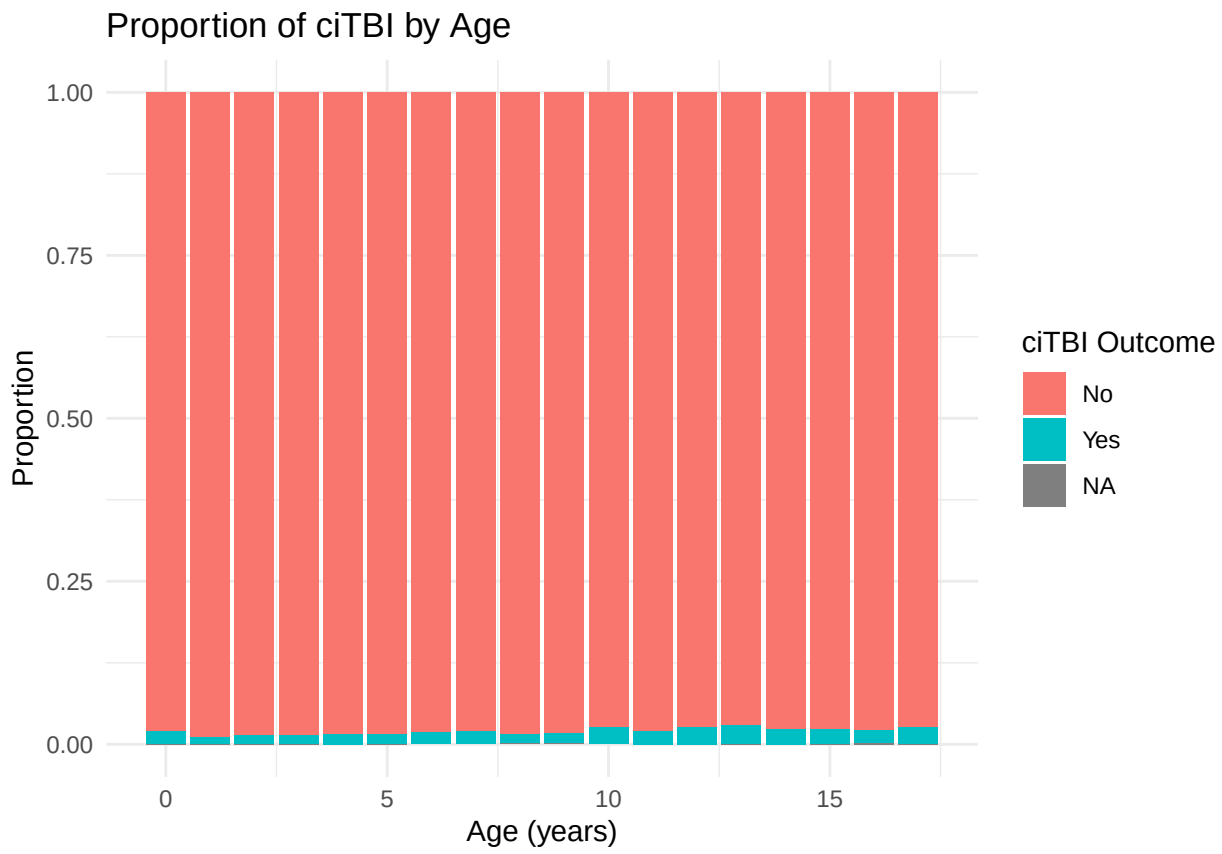


8

```
citbi_clean <- citbi_clean %>%
  mutate(age_years = floor(age_months / 12))
```

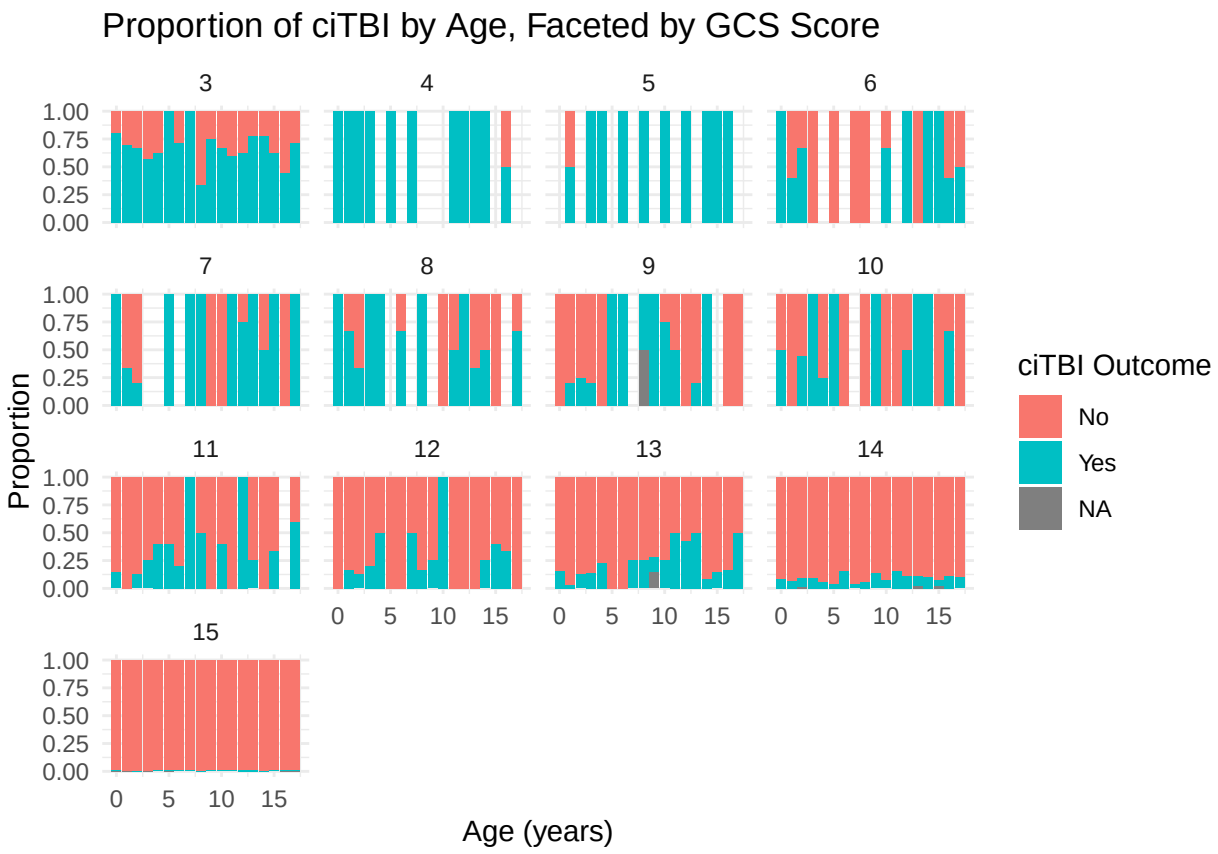
(a) Stacked normalized bar chart: proportion of ciTBI by age

```
ggplot(citbi_clean, aes(x = age_years, fill = citbi_outcome)) +
  geom_bar(position = "fill") +
  labs(
    title = "Proportion of ciTBI by Age",
    x = "Age (years)",
    y = "Proportion",
    fill = "ciTBI Outcome"
  ) +
  theme_minimal()
```



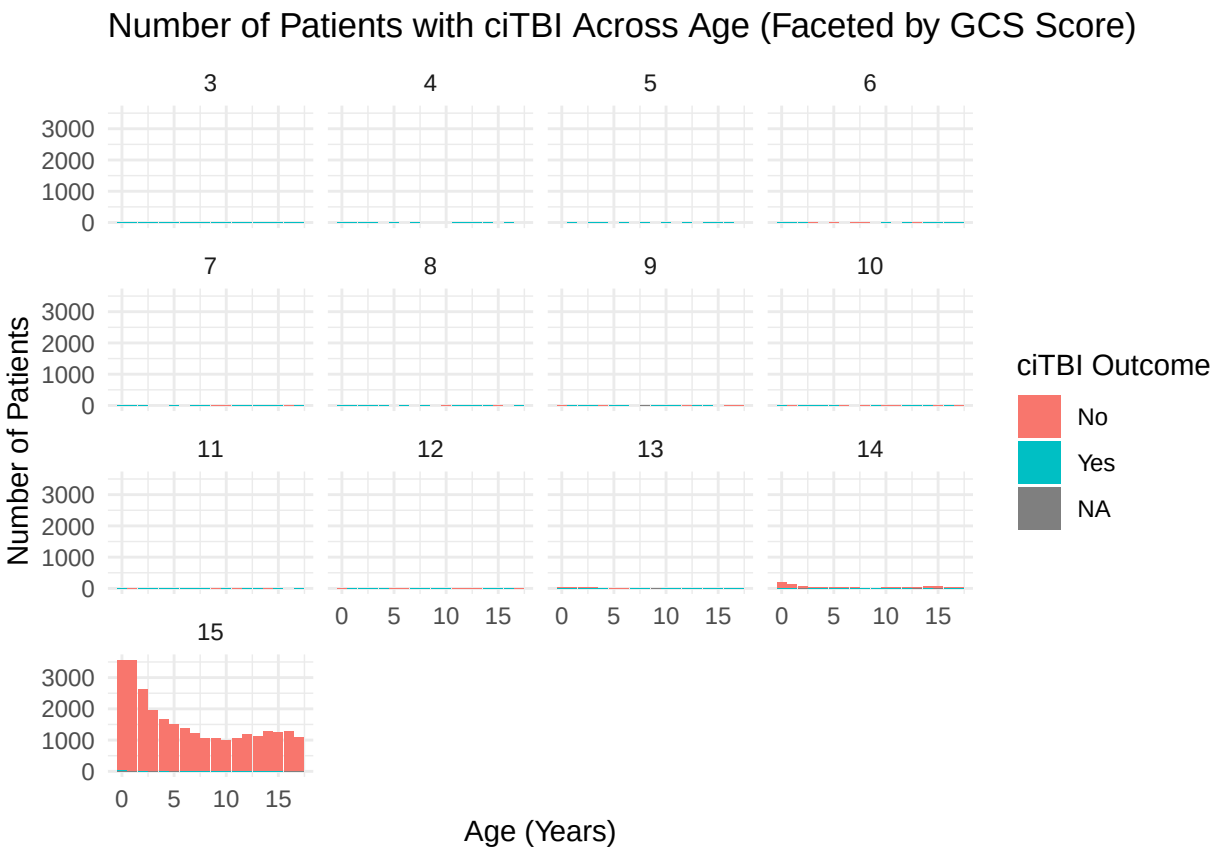
(b) Stacked normalized bar chart faceted by GCS total)

```r
citbi_clean <- citbi_clean %>%
  mutate(age_years = floor(age_months / 12))
ggplot(citbi_clean, aes(x = age_years, fill = citbi_outcome)) +
  geom_bar(position = "fill") +
  facet_wrap(~ gcs_total) +
  labs(
    title = "Proportion of ciTBI by Age, Faceted by GCS Score",
    x = "Age (years)",
    y = "Proportion",
    fill = "ciTBI Outcome"
  ) +
  theme_minimal()
```



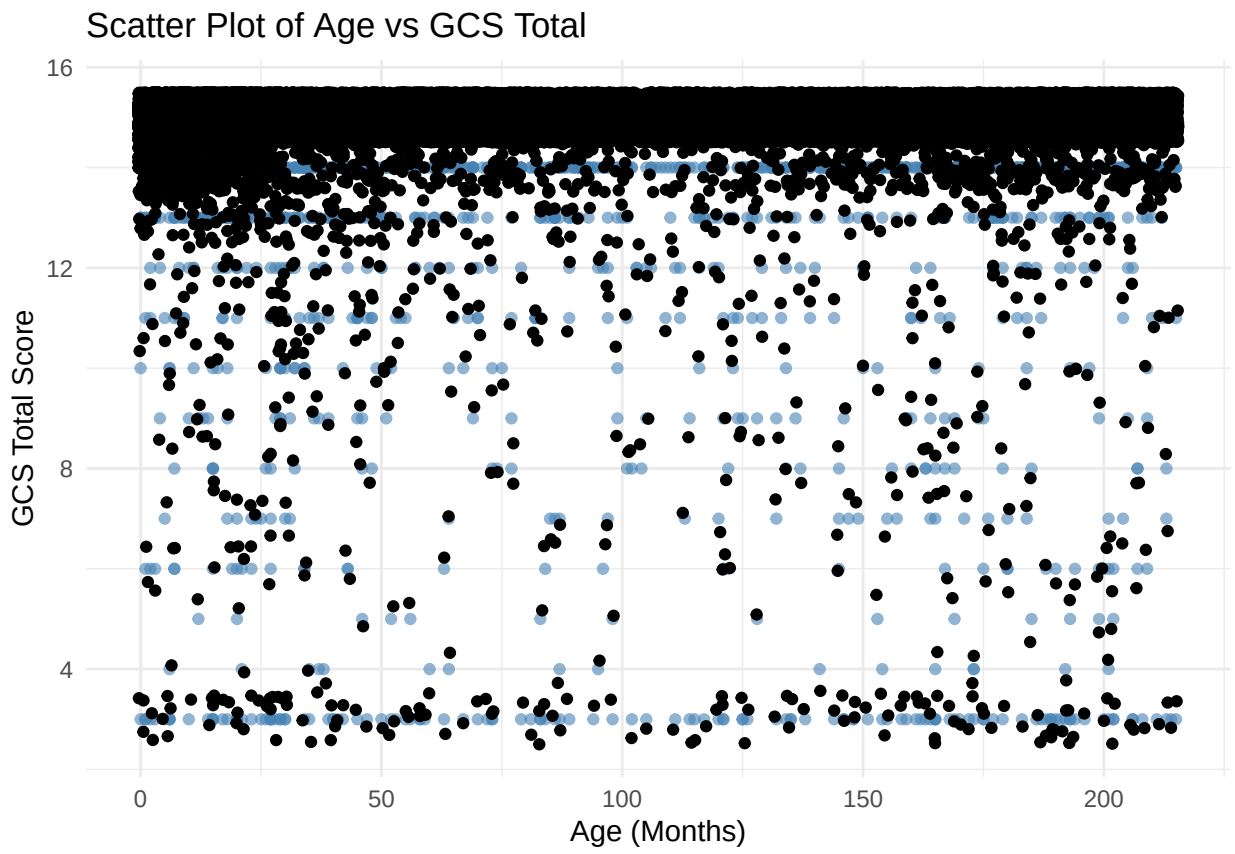
(c) Stacked bar chart of counts faceted by GCS total

```r
ggplot(citbi_clean, aes(x = age_years, fill = citbi_outcome)) +
  geom_bar(position = "stack") +
  facet_wrap(~ gcs_total) +
  labs(
    title = "Number of Patients with ciTBI Across Age (Faceted by GCS Score)",
    x = "Age (Years)",
    y = "Number of Patients",
    fill = "ciTBI Outcome"
  ) +
  theme_minimal()
```



9

```
ggplot(citbi_clean, aes(x = age_months, y = gcs_total)) +
  geom_point(alpha = 0.6, color = "steelblue") +
  geom_jitter(width = 0.5, height = 0.5) +  # avoid overplotting
  labs(
    title = "Scatter Plot of Age vs GCS Total",
    x = "Age (Months)",
    y = "GCS Total Score"
  ) +
  theme_minimal()
```



10

```
summary_stats <- citbi_clean %>%
  group_by(loc_duration, citbi_outcome) %>%
  summarise(
    mean_gcs = mean(gcs_total, na.rm = TRUE),
    mean_age_months = mean(age_months, na.rm = TRUE),
    count = n()
  ) %>%
  arrange(loc_duration, citbi_outcome)
```

```
## `summarise()` has grouped output by 'loc_duration'. You can override using the
## `.groups` argument.
```

```
summary_stats
```

```
## # A tibble: 13 x 5
## # Groups:   loc_duration [5]
##    loc_duration citbi_outcome mean_gcs mean_age_months count
##    <ord>        <fct>            <dbl>           <dbl> <int>
##  1 <5 sec       No                14.9            130.    652
##  2 <5 sec       Yes               14.8             95.4     8
```

```
##  3 5 sec-<1 min No              14.9           122.   1283
##  4 5 sec-<1 min Yes             14.0           117.     25
##  5 1-5 min      No              14.8           128.    723
##  6 1-5 min      Yes             12.7           120.     30
##  7 1-5 min      <NA>            15             214       1
##  8 >5 min       No              13.8            96.0   168
##  9 >5 min       Yes              5.44          106.     90
## 10 >5 min       <NA>             9              99       1
## 11 <NA>         No              14.9            80.1 26993
## 12 <NA>         Yes             11.7            93.0   394
## 13 <NA>         <NA>            14.5           112.     11
```

‘‘‘

R Markdown