Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

Emma Mo and 906542365

Display machine information for reproducibility:

```
sessionInfo()
```

```
R version 4.4.1 (2024-06-14)
Platform: x86_64-apple-darwin20
Running under: macOS Monterey 12.4

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRlapack.dylib;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/New_York
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

loaded via a namespace (and not attached):
 [1] compiler_4.4.1    fastmap_1.2.0     cli_3.6.3         tools_4.4.1
 [5] htmltools_0.5.8.1 rstudioapi_0.16.0 yaml_2.3.10       rmarkdown_2.29
 [9] knitr_1.49        jsonlite_1.8.9    xfun_0.50         digest_0.6.37
[13] rlang_1.1.4       evaluate_1.0.3
```

Load necessary libraries (you can add more as needed).

```
library(arrow)
```

```
Attaching package: 'arrow'

The following object is masked from 'package:utils':

    timestamp
```

```
library(gtsummary)
library(memuse)
library(pryr)
```

```
Attaching package: 'pryr'

The following object is masked from 'package:gtsummary':

    where
```

```
library(R.utils)
```

```
Loading required package: R.oo

Loading required package: R.methodsS3

R.methodsS3 v1.8.2 (2022-06-13 22:00:14 UTC) successfully loaded. See ?R.methodsS3 for help.

R.oo v1.27.0 (2024-11-01 18:00:02 UTC) successfully loaded. See ?R.oo for help.

Attaching package: 'R.oo'

The following object is masked from 'package:R.methodsS3':

    throw
```

```
The following objects are masked from 'package:methods':

    getClasses, getMethods

The following objects are masked from 'package:base':

    attach, detach, load, save

R.utils v2.12.3 (2023-11-18 01:00:02 UTC) successfully loaded. See ?R.utils for help.


Attaching package: 'R.utils'

The following object is masked from 'package:arrow':

    timestamp

The following object is masked from 'package:utils':

    timestamp

The following objects are masked from 'package:base':

    cat, commandArgs, getOption, isOpen, nullfile, parse, use, warnings
```

```
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 purrr::compose()      masks pryr::compose()
x lubridate::duration() masks arrow::duration()
x tidyr::extract()      masks R.utils::extract()
x dplyr::filter()       masks stats::filter()
```

```
x dplyr::lag()          masks stats::lag()
x purrr::partial()      masks pryr::partial()
x dplyr::where()        masks pryr::where(), gtsummary::where()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

Display your machine memory.

```
memuse::Sys.meminfo()
```

```
Totalram:  16.000 GiB
Freeram:    5.112 GiB
```

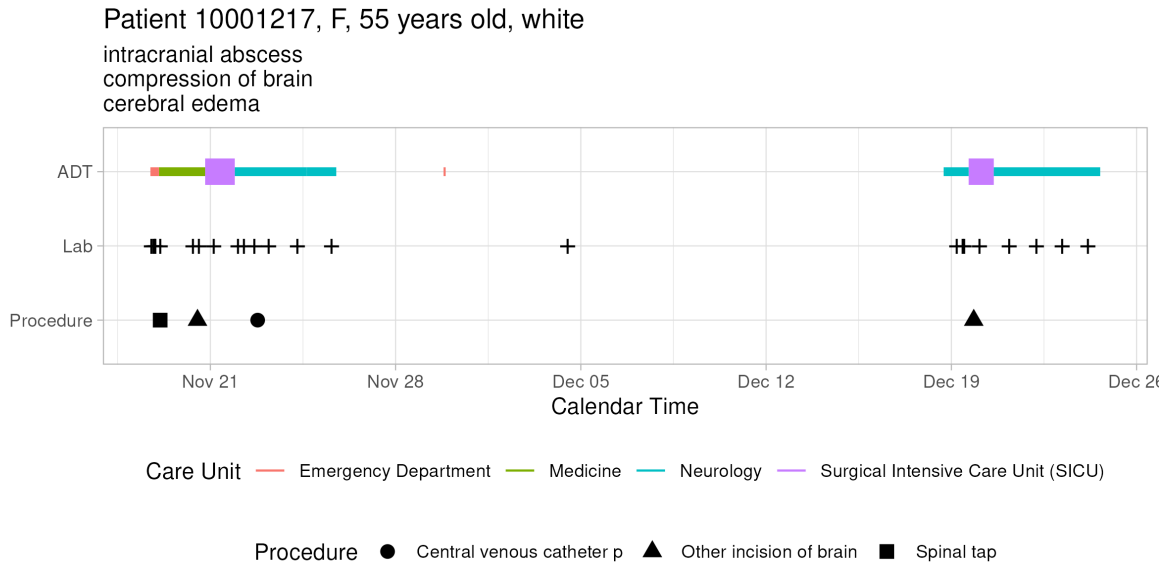
In this exercise, we use tidyverse (ggplot2, dplyr, etc) to explore the MIMIC-IV data introduced in homework 1 and to build a cohort of ICU stays.

Q1. Visualizing patient trajectory

Visualizing a patient's encounters in a health care system is a common task in clinical data analysis. In this question, we will visualize a patient's ADT (admission-discharge-transfer) history and ICU vitals in the MIMIC-IV data.

Q1.1 ADT history

A patient's ADT history records the time of admission, discharge, and transfer in the hospital. This figure shows the ADT history of the patient with `subject_id` 10001217 in the MIMIC-IV data. The x-axis is the calendar time, and the y-axis is the type of event (ADT, lab, procedure). The color of the line segment represents the care unit. The size of the line segment represents whether the care unit is an ICU/CCU. The crosses represent lab events, and the shape of the dots represents the type of procedure. The title of the figure shows the patient's demographic information and the subtitle shows top 3 diagnoses.



Do a similar visualization for the patient with `subject_id` 10063848 using ggplot.

Hint: We need to pull information from data files `patients.csv.gz`, `admissions.csv.gz`, `transfers.csv.gz`, `labevents.csv.gz`, `procedures_icd.csv.gz`, `diagnoses_icd.csv.gz`, `d_icd_procedures.csv.gz`, and `d_icd_diagnoses.csv.gz`. For the big file `labevents.csv.gz`, use the Parquet format you generated in Homework 2. For reproducibility, make the Parquet folder `labevents_pq` available at the current working directory `hw3`, for example, by a symbolic link. Make your code reproducible.

```
source_dir <- "../hw2/labevents.parquet"
target_dir <- "labevents_pq"

if (file.exists(target_dir) || dir.exists(target_dir)) {
  unlink(target_dir, recursive = TRUE)
}

file.symlink(from = source_dir, to = target_dir)
```

```
[1] TRUE
```

```
file.exists(target_dir)
```

```
[1] TRUE
```

```
parquet_dir <- "labevents_pq/"
labevents <- open_dataset(parquet_dir)
```

```
hosp_dir <- "~/mimic/hosp"

patients       <- read_csv(file.path(hosp_dir, "patients.csv.gz"))
```

```
Rows: 364627 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): gender, anchor_year_group
dbl  (3): subject_id, anchor_age, anchor_year
date (1): dod

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.
```

```
admissions     <- read_csv(file.path(hosp_dir, "admissions.csv.gz"))
```

```
Rows: 546028 Columns: 16
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (8): admission_type, admit_provider_id, admission_location, discharge_l...
dbl  (3): subject_id, hadm_id, hospital_expire_flag
dttm (5): admittime, dischtime, deathtime, edregtime, edouttime

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.
```

```
transfers      <- read_csv(file.path(hosp_dir, "transfers.csv.gz"))
```

```
Rows: 2413581 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): eventtype, careunit
dbl  (3): subject_id, hadm_id, transfer_id
dttm (2): intime, outtime

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.
```

```
procedures     <- read_csv(file.path(hosp_dir, "procedures_icd.csv.gz"))
```

```
Rows: 859655 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): icd_code
dbl  (4): subject_id, hadm_id, seq_num, icd_version
date (1): chartdate

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.
```

```
diagnoses      <- read_csv(file.path(hosp_dir, "diagnoses_icd.csv.gz"))
```

```
Rows: 6364488 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): icd_code
dbl (4): subject_id, hadm_id, seq_num, icd_version

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.
```

```
d_icd_procedures <- read_csv(file.path(hosp_dir, "d_icd_procedures.csv.gz"))
```

```
Rows: 86423 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (2): icd_code, long_title
dbl (1): icd_version

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.
```

```
d_icd_diagnoses  <- read_csv(file.path(hosp_dir, "d_icd_diagnoses.csv.gz"))
```

```
Rows: 112107 Columns: 3
-- Column specification --------------------------------------------------------
Delimiter: ","
```

```
chr (2): icd_code, long_title
dbl (1): icd_version

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
subject_id <- 10063848

# Filter patient info
patient_info <- patients %>%
  filter(subject_id == !!subject_id)

# Filter admissions info
admissions_info <- admissions %>%
  filter(subject_id == !!subject_id)

# Process transfers (ADT) info, converting to Date only
transfers_info <- transfers %>%
  filter(subject_id == !!subject_id) %>%
  mutate(
    event_type = "ADT",
    start_date = as.Date(intime),
    end_date   = as.Date(outtime),
    is_icu     = str_detect(careunit, "ICU|CCU")
  )

# Process lab events, converting charttime to Date
lab_info <- labevents %>%
  filter(subject_id == !!subject_id) %>%
  collect() %>%
  mutate(
    event_type = "Lab",
    start_date = as.Date(charttime)
  )

# Process procedures, converting chartdate to Date
procedures_info <- procedures %>%
  filter(subject_id == !!subject_id) %>%
  left_join(d_icd_procedures, by = "icd_code") %>%
  mutate(
    event_type = "Procedure",
    start_date = as.Date(chartdate)
```

```
  )

# Get top 3 diagnoses
diagnoses_info <- diagnoses %>%
  filter(subject_id == !!subject_id) %>%
  left_join(d_icd_diagnoses, by = "icd_code") %>%
  select(subject_id, hadm_id, icd_code, long_title)
```

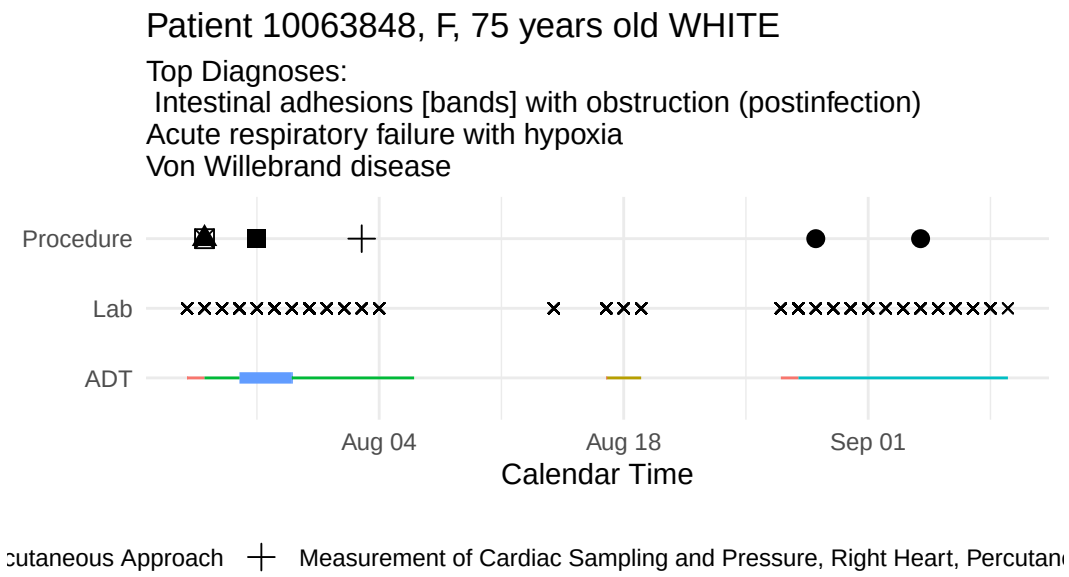
```
Warning in left_join(., d_icd_diagnoses, by = "icd_code"): Detected an unexpected many-to-ma
i Row 17 of `x` matches multiple rows in `y`.
i Row 15793 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

```
top_diagnoses <- diagnoses_info %>%
  slice_head(n = 3) %>%
  pull(long_title)
```

```
# Convert is_icu to a factor (for proper mapping)
transfers_info <- transfers_info %>%
  mutate(is_icu = as.factor(is_icu))

# Create the plot using ggplot2 with Date-based x-axis
ggplot() +
  # ADT segments
  geom_segment(
    data = transfers_info,
    aes(x = start_date, xend = end_date,
        y = event_type, yend = event_type,
        color = careunit, linewidth = is_icu)
  ) +
  scale_linewidth_manual(values = c("FALSE" = 0.5, "TRUE" = 2)) +

  # Lab events as crosses
  geom_point(
    data = lab_info,
    aes(x = start_date, y = event_type),
    shape = 4  # cross shape
  ) +

  # Procedures as points with different shapes
```

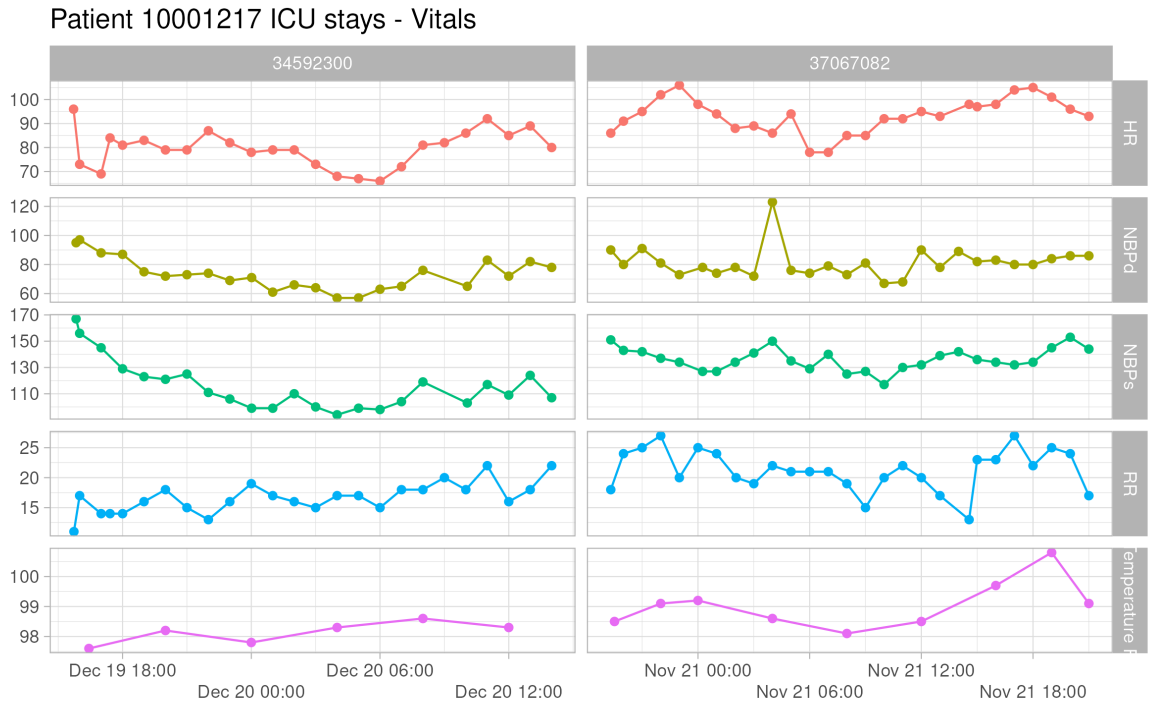
```
  geom_point(
    data = procedures_info,
    aes(x = start_date, y = event_type, shape = long_title),
    size = 3
  ) +

  labs(
    title = paste0(
      "Patient ", subject_id, ", ",
      patient_info$gender, ", ",
      patient_info$anchor_age, " years old ",
      admissions_info$race
    ),
    subtitle = paste("Top Diagnoses:\n", paste(top_diagnoses, collapse = "\n")),
    x = "Calendar Time",
    y = NULL,
    color = "Care Unit",
    shape = "Procedure",
    linewidth = "ICU/CCU"
  ) +
  theme_minimal() +
  theme(legend.position = "bottom") +
  scale_x_date(date_breaks = "2 weeks", date_labels = "%b %d")
```

```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_segment()`).
```



Q1.2 ICU stays

ICU stays are a subset of ADT history. This figure shows the vitals of the patient `10001217` during ICU stays. The x-axis is the calendar time, and the y-axis is the value of the vital. The color of the line represents the type of vital. The facet grid shows the abbreviation of the vital and the stay ID.



Do a similar visualization for the patient 10063848.

```
csv_file <- "~/mimic/icu/chartevents.csv.gz"
parquet_file <- "~/mimic/icu/csv_file"

parquet_data <- open_dataset(parquet_file, format = "parquet")

chartevents_data <- parquet_data %>%
  filter(subject_id %in% c(10063848))
chartevents <- collect(chartevents_data)
```

```
d_items     <- read_csv("~/mimic/icu/d_items.csv.gz")
```

```
Rows: 4095 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (6): label, abbreviation, linksto, category, unitname, param_type
dbl (3): itemid, lownormalvalue, highnormalvalue

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
chartevents_joined <- chartevents %>%
  left_join(d_items, by = "itemid")

subject_of_interest <- 10063848

vitals <- chartevents_joined %>%
  filter(
    subject_id == subject_of_interest,
    label %in% c("Heart Rate",
                 "Non Invasive Blood Pressure systolic",
                 "Non Invasive Blood Pressure diastolic",
                 "Respiratory Rate",
                 "Temperature Fahrenheit")
  ) %>%
  mutate(
    charttime = ymd_hms(charttime)  # parse as POSIXct
  )
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `charttime = ymd_hms(charttime)`.
Caused by warning:
!  8 failed to parse.
```

```r
icustays <- read_csv("~/mimic/icu/icustays.csv.gz") %>%
  filter(subject_id == subject_of_interest) %>%
  mutate(
    intime  = ymd_hms(intime),
    outtime = ymd_hms(outtime)
  )
```

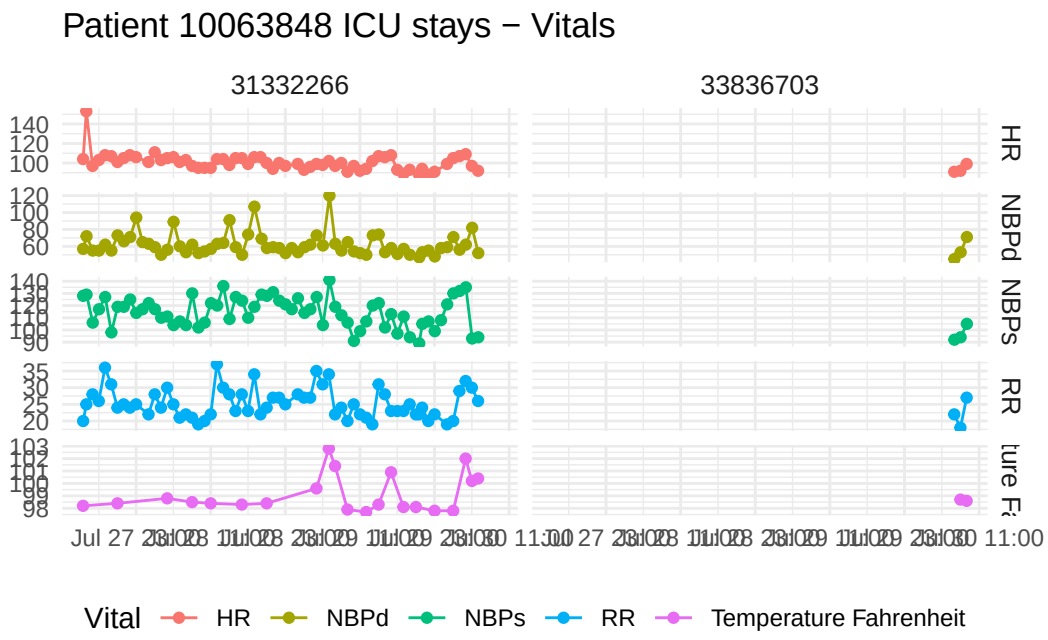
```
Rows: 94458 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): first_careunit, last_careunit
dbl  (4): subject_id, hadm_id, stay_id, los
dttm (2): intime, outtime

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
vitals_icu <- vitals %>%
  inner_join(icustays, by = c("subject_id", "stay_id"))
```

```r
vitals_icu <- vitals_icu %>%
  mutate(
    vital_abbr = case_when(
      label == "Heart Rate" ~ "HR",
      label == "Non Invasive Blood Pressure systolic"  ~ "NBPs",
      label == "Non Invasive Blood Pressure diastolic" ~ "NBPd",
      label == "Respiratory Rate" ~ "RR",
      label == "Temperature Fahrenheit" ~ "Temperature Fahrenheit",
      TRUE ~ label
    )
  )
```

```r
ggplot(vitals_icu, aes(x = charttime, y = valuenum, color = vital_abbr)) +
  geom_line() +
  geom_point() +
  facet_grid(vital_abbr ~ stay_id, scales = "free_y") +
  scale_x_datetime(date_breaks = "12 hours", date_labels = "%b %d %H:%M") +
  labs(
    title = paste("Patient", subject_of_interest, "ICU stays - Vitals"),
    x = NULL,
    y = NULL,  # or "Value"
    color = "Vital"
  ) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    strip.text = element_text(size = 10)
  )
```

```
Warning: Removed 8 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Q2. ICU stays

`icustays.csv.gz` (https://mimic.mit.edu/docs/iv/modules/icu/icustays/) contains data about Intensive Care Units (ICU) stays. The first 10 lines are

```
zcat < ~/mimic/icu/icustays.csv.gz | head
```

```
subject_id,hadm_id,stay_id,first_careunit,last_careunit,intime,outtime,los
10000032,29079034,39553978,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit (MI
10000690,25860671,37081114,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit (MI
10000980,26913865,39765666,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit (MI
10001217,24597018,37067082,Surgical Intensive Care Unit (SICU),Surgical Intensive Care Unit
10001217,27703517,34592300,Surgical Intensive Care Unit (SICU),Surgical Intensive Care Unit
10001725,25563031,31205490,Medical/Surgical Intensive Care Unit (MICU/SICU),Medical/Surgical
10001843,26133978,39698942,Medical/Surgical Intensive Care Unit (MICU/SICU),Medical/Surgical
10001884,26184834,37510196,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit (MI
10002013,23581541,39060235,Cardiac Vascular Intensive Care Unit (CVICU),Cardiac Vascular Inte
```

Q2.1 Ingestion

Import `icustays.csv.gz` as a tibble `icustays_tble`.

```
icustays_tble <- read_csv("~/mimic/icu/icustays.csv.gz")
```

```
Rows: 94458 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): first_careunit, last_careunit
dbl  (4): subject_id, hadm_id, stay_id, los
dttm (2): intime, outtime

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.
```

Q2.2 Summary and visualization

How many unique `subject_id`? Can a `subject_id` have multiple ICU stays? Summarize the number of ICU stays per `subject_id` by graphs.

```
num_unique_subjects <- icustays_tble %>%
  summarize(n_unique_subjects = n_distinct(subject_id)) %>%
  pull(n_unique_subjects)
num_unique_subjects
```

```
[1] 65366
```

There are 65366 unique `subject_id`.

```
subject_stays <- icustays_tble %>%
  group_by(subject_id) %>%
  summarize(n_stays = n(), .groups = "drop")
subject_stays
```

```
# A tibble: 65,366 x 2
   subject_id n_stays
        <dbl>   <int>
 1   10000032       1
 2   10000690       1
 3   10000980       1
 4   10001217       2
 5   10001725       1
```

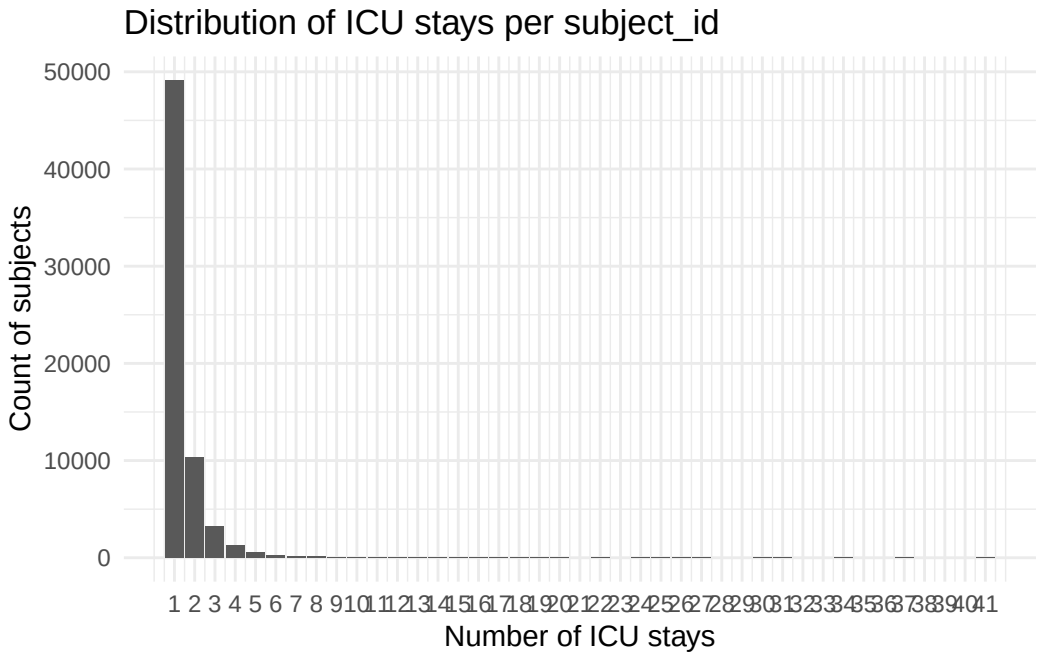
```
 6   10001843       1
 7   10001884       1
 8   10002013       1
 9   10002114       1
10   10002155       3
# i 65,356 more rows
```

```r
sum(subject_stays$n_stays > 1)
```

```
[1] 16242
```

Yes, a `subject_id` can have multiple ICU stays.

```r
ggplot(subject_stays, aes(x = n_stays)) +
  geom_bar() +
  scale_x_continuous(breaks = 1:max(subject_stays$n_stays)) +
  labs(
    title = "Distribution of ICU stays per subject_id",
    x = "Number of ICU stays",
    y = "Count of subjects"
  ) +
  theme_minimal()
```



Q3. `admissions` data

Information of the patients admitted into hospital is available in `admissions.csv.gz`. See https://mimic.mit.edu/docs/iv/modules/hosp/admissions/ for details of each field in this file. The first 10 lines are

```
zcat < ~/mimic/hosp/admissions.csv.gz | head
```

```
subject_id,hadm_id,admittime,dischtime,deathtime,admission_type,admit_provider_id,admission_l
10000032,22595853,2180-05-06 22:23:00,2180-05-07 17:15:00,,URGENT,P49AFC,TRANSFER FROM HOSPI
10000032,22841357,2180-06-26 18:27:00,2180-06-27 18:49:00,,EW EMER.,P784FA,EMERGENCY ROOM,HO
10000032,25742920,2180-08-05 23:44:00,2180-08-07 17:50:00,,EW EMER.,P19UTS,EMERGENCY ROOM,HOS
10000032,29079034,2180-07-23 12:35:00,2180-07-25 17:55:00,,EW EMER.,P06OTX,EMERGENCY ROOM,HO
10000068,25022803,2160-03-03 23:16:00,2160-03-04 06:26:00,,EU OBSERVATION,P39NWO,EMERGENCY RO
10000084,23052089,2160-11-21 01:56:00,2160-11-25 14:52:00,,EW EMER.,P42H7G,WALK-IN/SELF REFE
10000084,29888819,2160-12-28 05:11:00,2160-12-28 16:07:00,,EU OBSERVATION,P35NE4,PHYSICIAN RE
10000108,27250926,2163-09-27 23:17:00,2163-09-28 09:04:00,,EU OBSERVATION,P40JML,EMERGENCY RO
10000117,22927623,2181-11-15 02:05:00,2181-11-15 14:52:00,,EU OBSERVATION,P47EY8,EMERGENCY RO
```

Q3.1 Ingestion

Import `admissions.csv.gz` as a tibble `admissions_tble`.

```
admissions_tble <- read_csv("~/mimic/hosp/admissions.csv.gz")
```

```
Rows: 546028 Columns: 16
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (8): admission_type, admit_provider_id, admission_location, discharge_l...
dbl  (3): subject_id, hadm_id, hospital_expire_flag
dttm (5): admittime, dischtime, deathtime, edregtime, edouttime

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.
```

Q3.2 Summary and visualization

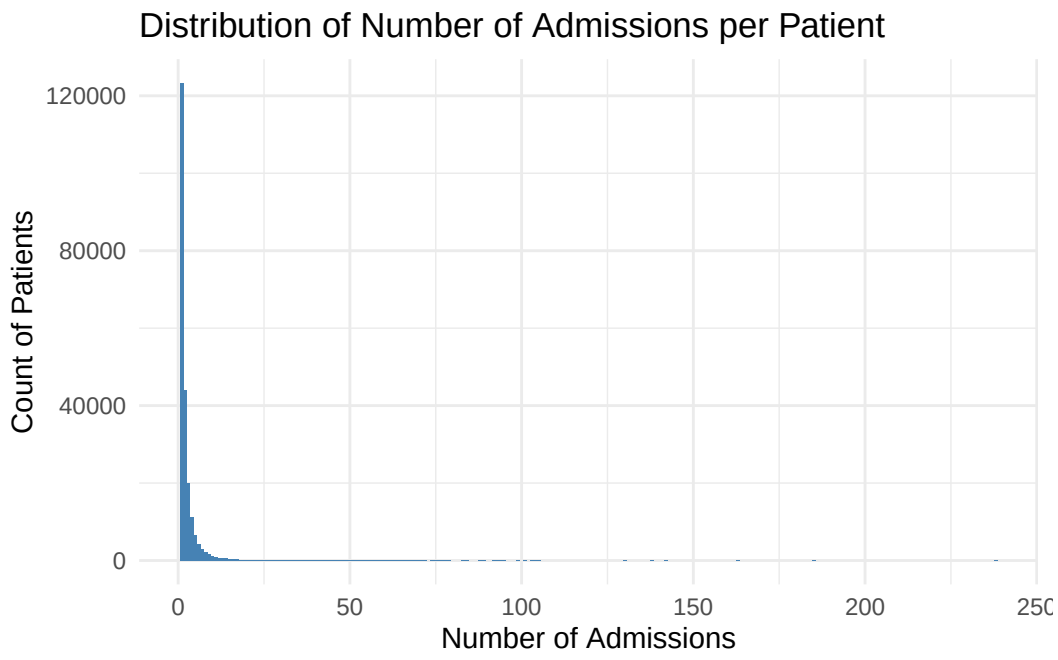
Summarize the following information by graphics and explain any patterns you see.

- number of admissions per patient

- admission hour (anything unusual?)
- admission minute (anything unusual?)
- length of hospital stay (from admission to discharge) (anything unusual?)

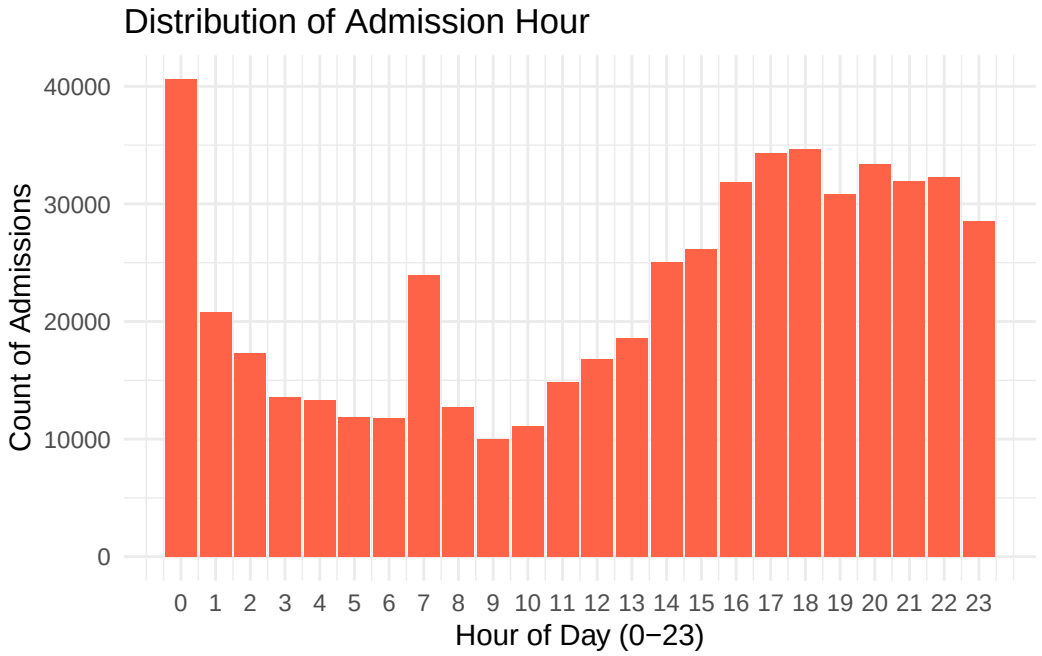
According to the MIMIC-IV documentation,

```
library(tidyverse)
library(lubridate)  # For working with dates/times

adm_per_patient <- admissions_tble %>%
  group_by(subject_id) %>%
  summarize(n_admissions = n(), .groups = "drop")

ggplot(adm_per_patient, aes(x = n_admissions)) +
  geom_bar(fill = "steelblue") +
  labs(
    title = "Distribution of Number of Admissions per Patient",
    x = "Number of Admissions",
    y = "Count of Patients"
  ) +
  theme_minimal()
```



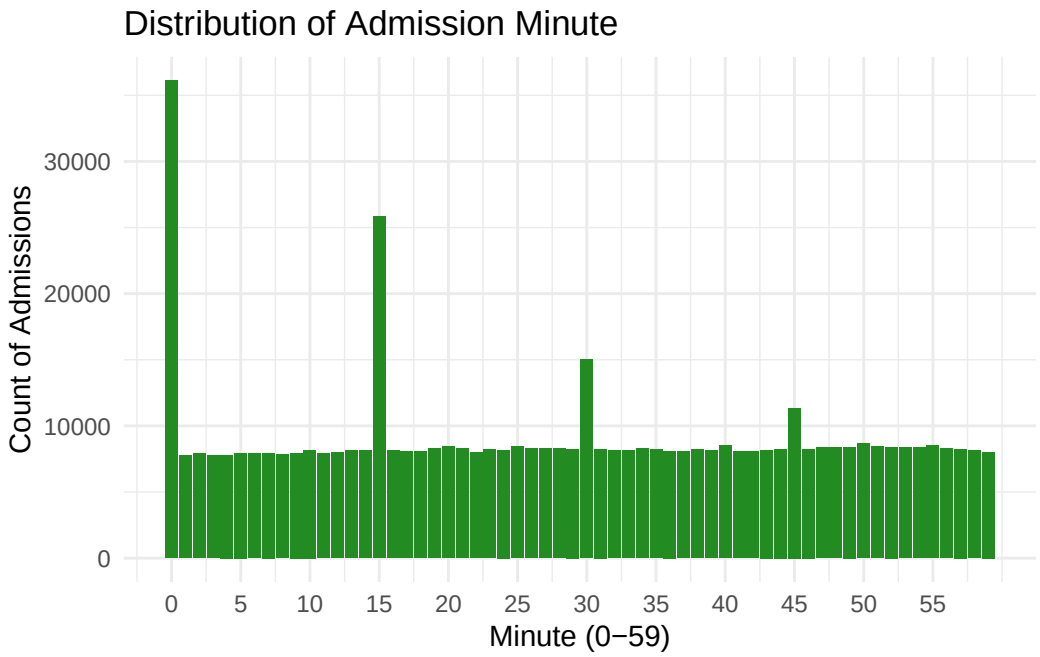
- Many patients typically have only 1 admission (common in hospital databases).
- A long, thin tail extending out to higher numbers of admissions, meaning fewer patients have many admissions.

```r
admissions_tble <- admissions_tble %>%
  mutate(admission_hour = hour(admittime))

ggplot(admissions_tble, aes(x = admission_hour)) +
  geom_bar(fill = "tomato") +
  scale_x_continuous(breaks = 0:23) +
  labs(
    title = "Distribution of Admission Hour",
    x = "Hour of Day (0-23)",
    y = "Count of Admissions"
  ) +
  theme_minimal()
```



- A very high bar at hour = 0 (midnight).
- Fewer admissions during the early morning hours (1–9 AM), with another spike at 7 AM.
- A gradual rise and plateau from 10 AM through evening hours.

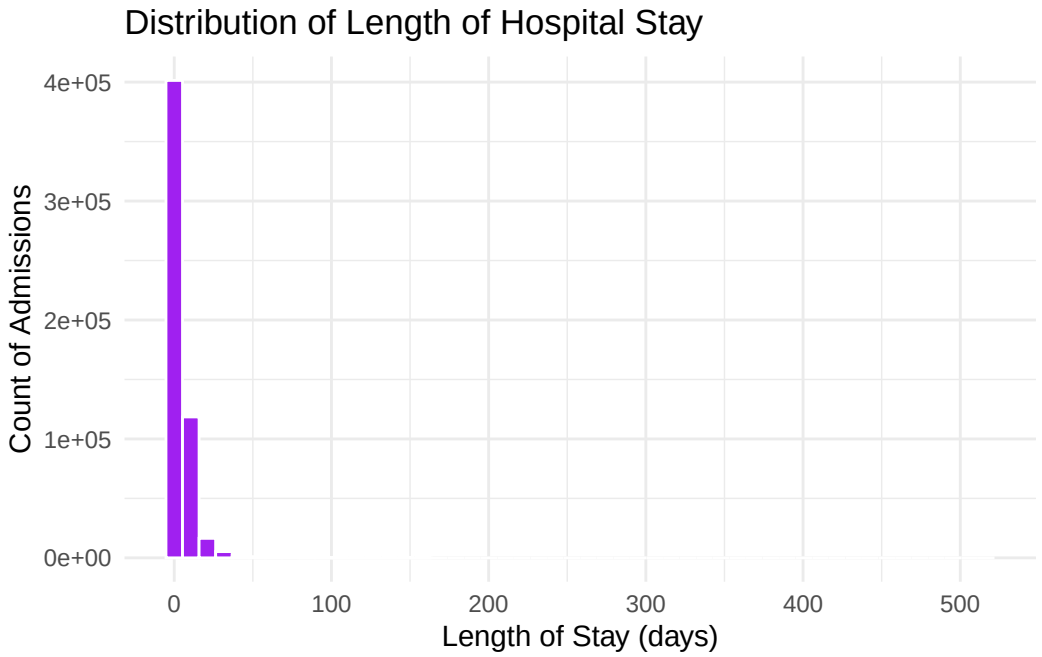
```r
# 1. Extract the minute of admission
admissions_tble <- admissions_tble %>%
  mutate(admission_minute = minute(admittime))

# 2. Plot the distribution of admission_minute
ggplot(admissions_tble, aes(x = admission_minute)) +
  geom_bar(fill = "forestgreen") +
  scale_x_continuous(breaks = seq(0, 59, by = 5)) +
  labs(
    title = "Distribution of Admission Minute",
    x = "Minute (0-59)",
    y = "Count of Admissions"
  ) +
  theme_minimal()
```



Every 15 minutes there is a spike. This is likely due to rounding of the admission time to the nearest 15 minutes.

```r
admissions_tble <- admissions_tble %>%
  mutate(
    length_of_stay_days = as.numeric(difftime(dischtime, admittime, units = "days"))
  )
```

```r
ggplot(admissions_tble, aes(x = length_of_stay_days)) +
  geom_histogram(bins = 50, fill = "purple", color = "white") +
  labs(
    title = "Distribution of Length of Hospital Stay",
    x = "Length of Stay (days)",
    y = "Count of Admissions"
  ) +
  theme_minimal()
```



- A large number of very short stays (near 0–2 days).
- The distribution tapers off quickly but extends to very long stays (hundreds of days).
- A classic right-skewed distribution: many short stays, fewer very long ones.

All dates in the database have been shifted to protect patient confidentiality. Dates will be internally consistent for the same patient, but randomly distributed in the future. Dates of birth which occur in the present time are not true dates of birth. Furthermore, dates of birth which occur before the year 1900 occur if the patient is older than 89. In these cases, the patient's age at their first admission has been fixed to 300.

Q4. `patients` data

Patient information is available in `patients.csv.gz`. See https://mimic.mit.edu/docs/iv/modules/hosp/patients/ for details of each field in this file. The first 10 lines are

```
zcat < ~/mimic/hosp/patients.csv.gz | head
```

```
subject_id,gender,anchor_age,anchor_year,anchor_year_group,dod
10000032,F,52,2180,2014 - 2016,2180-09-09
10000048,F,23,2126,2008 - 2010,
10000058,F,33,2168,2020 - 2022,
10000068,F,19,2160,2008 - 2010,
10000084,M,72,2160,2017 - 2019,2161-02-13
10000102,F,27,2136,2008 - 2010,
10000108,M,25,2163,2014 - 2016,
10000115,M,24,2154,2017 - 2019,
10000117,F,48,2174,2008 - 2010,
```

Q4.1 Ingestion

Import `patients.csv.gz` (https://mimic.mit.edu/docs/iv/modules/hosp/patients/) as a tibble `patients_tble`.

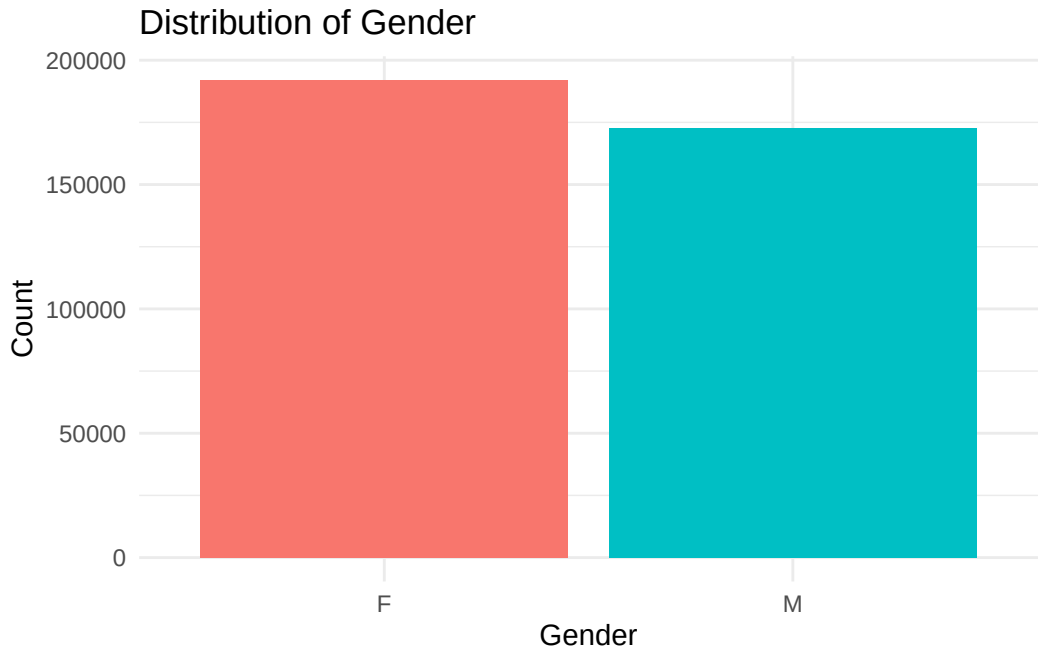
```
patients_tble <- read_csv("~/mimic/hosp/patients.csv.gz")
```

```
Rows: 364627 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): gender, anchor_year_group
dbl  (3): subject_id, anchor_age, anchor_year
date (1): dod

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.
```

Q4.2 Summary and visualization

Summarize variables `gender` and `anchor_age` by graphics, and explain any patterns you see.

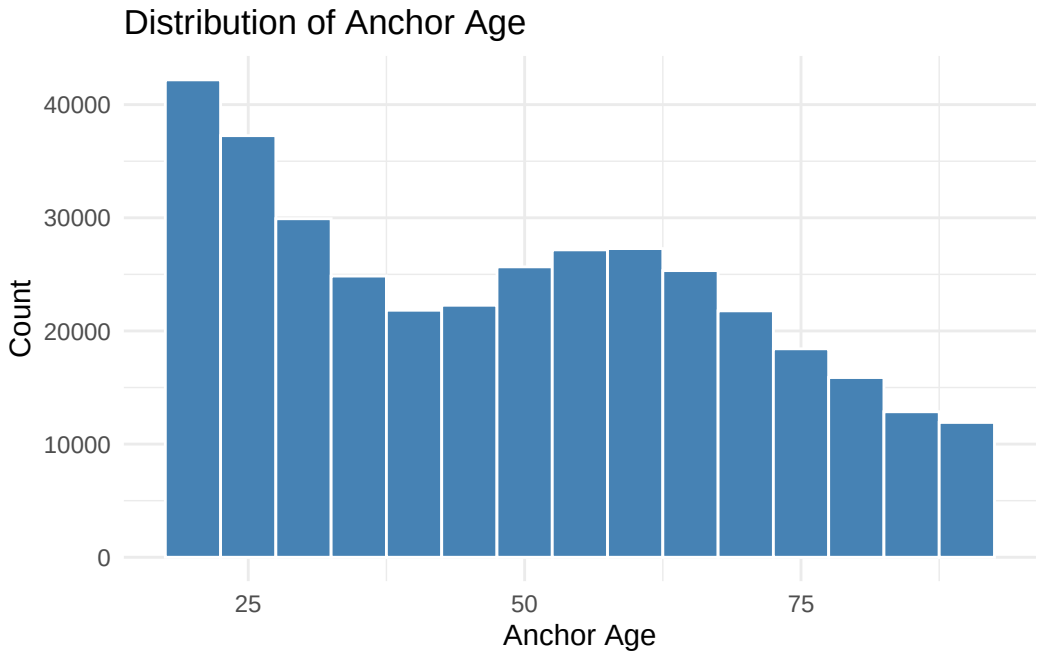
```r
library(tidyverse)

gender_count <- patients_tble %>%
  group_by(gender) %>%
  summarize(count = n(), .groups = "drop")

ggplot(gender_count, aes(x = gender, y = count, fill = gender)) +
  geom_bar(stat = "identity") +
  labs(
    title = "Distribution of Gender",
    x = "Gender",
    y = "Count"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



Female is slightly more than male.

```r
ggplot(patients_tble, aes(x = anchor_age)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(
    title = "Distribution of Anchor Age",
```

```
    x = "Anchor Age",
    y = "Count"
  ) +
  theme_minimal()
```



- The majority of patients cluster around young adult ages (e.g., 20s–30s).
- There is also a substantial number of middle-aged and older adults.

Q5. Lab results

`labevents.csv.gz` (https://mimic.mit.edu/docs/iv/modules/hosp/labevents/) contains all laboratory measurements for patients. The first 10 lines are

```
zcat < ~/mimic/hosp/labevents.csv.gz | head
```

```
labevent_id,subject_id,hadm_id,specimen_id,itemid,order_provider_id,charttime,storetime,valu
1,10000032,,2704548,50931,P69FQC,2180-03-23 11:51:00,2180-03-23 15:56:00,___,95,mg/dL,70,100
2,10000032,,36092842,51071,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,
3,10000032,,36092842,51074,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,
4,10000032,,36092842,51075,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,"
5,10000032,,36092842,51079,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,
```

```
6,10000032,,36092842,51087,P69FQC,2180-03-23 11:51:00,,,,,,,ROUTINE,RANDOM.
7,10000032,,36092842,51089,P69FQC,2180-03-23 11:51:00,2180-03-23 16:15:00,,,,,,ROUTINE,PRESU
8,10000032,,36092842,51090,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,MI
9,10000032,,36092842,51092,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,"(
```

`d_labitems.csv.gz` (https://mimic.mit.edu/docs/iv/modules/hosp/d_labitems/) is the dictionary of lab measurements.

```
zcat < ~/mimic/hosp/d_labitems.csv.gz | head
```

```
itemid,label,fluid,category
50801,Alveolar-arterial Gradient,Blood,Blood Gas
50802,Base Excess,Blood,Blood Gas
50803,"Calculated Bicarbonate, Whole Blood",Blood,Blood Gas
50804,Calculated Total CO2,Blood,Blood Gas
50805,Carboxyhemoglobin,Blood,Blood Gas
50806,"Chloride, Whole Blood",Blood,Blood Gas
50808,Free Calcium,Blood,Blood Gas
50809,Glucose,Blood,Blood Gas
50810,"Hematocrit, Calculated",Blood,Blood Gas
```

We are interested in the lab measurements of creatinine (50912), potassium (50971), sodium (50983), chloride (50902), bicarbonate (50882), hematocrit (51221), white blood cell count (51301), and glucose (50931). Retrieve a subset of `labevents.csv.gz` that only containing these items for the patients in `icustays_tble`. Further restrict to the last available measurement (by `storetime`) before the ICU stay. The final `labevents_tble` should have one row per ICU stay and columns for each lab measurement.

```
> labevents_tble
# A tibble: 88,086 × 10
   subject_id  stay_id bicarbonate chloride creatinine glucose potassium sodium hematocrit   wbc
        <dbl>    <dbl>       <dbl>    <dbl>      <dbl>   <dbl>     <dbl>  <dbl>      <dbl> <dbl>
 1   10000032 39553978          25       95        0.7     102       6.7    126       41.1   6.9
 2   10000690 37081114          26      100          1      85       4.8    137       36.1   7.1
 3   10000980 39765666          21      109        2.3       89       3.9    144       27.3   5.3
 4   10001217 34592300          30      104        0.5       87       4.1    142       37.4   5.4
 5   10001217 37067082          22      108        0.6      112       4.2    142       38.1  15.7
 6   10001725 31205490          NA       98         NA       NA       4.1    139         NA    NA
 7   10001843 39698942          28       97        1.3      131       3.9    138       31.4  10.4
 8   10001884 37510196          30       88        1.1      141       4.5    130       39.7  12.2
 9   10002013 39060235          24      102        0.9      288       3.5    137       34.9   7.2
10   10002114 34672098          18       NA        3.1       95       6.5    125       34.3  16.8
# ℹ 88,076 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Hint: Use the Parquet format you generated in Homework 2. For reproducibility, make `labevents_pq` folder available at the current working directory `hw3`, for example, by a symbolic link.

```
library(arrow)

# Read the gzipped CSV file
labitems <- read_csv_arrow("~/mimic/hosp/d_labitems.csv.gz")

# Write the data to a Parquet file
write_parquet(labitems, "labitems_tble.parquet")

# Optionally, open the Parquet dataset as a table
labitems_tble <- open_dataset("labitems_tble.parquet", format = "parquet")
```

```
itemid_label <- c(
  "50912" = "creatinine",
  "50971" = "potassium",
  "50983" = "sodium",
  "50902" = "chloride",
  "50882" = "bicarbonate",
  "51221" = "hematocrit",
  "51301" = "white_blood_cells",
  "50931" = "glucose"
)
```

```
labevents_tble <- open_dataset("labevents_pq", format = "parquet") |>
  to_duckdb() |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in% c(50912, 50971, 50983, 50902, 50882, 51221, 51301, 50931)) |>
  left_join(
    select(icustays_tble, subject_id, stay_id, intime),
    by = c("subject_id"),
    copy = TRUE
  ) |>
  filter(storetime < intime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_max(storetime, n = 1) |>
  select(-storetime, -intime) |>
  ungroup() |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  rename_at(
```

```
    vars(names(itemid_label)),
    ~ itemid_label[.]
    ) |>
  show_query() |>
  collect() |>
  arrange(subject_id, stay_id) |>
  relocate(subject_id, stay_id)
```

```
<SQL>
SELECT
  subject_id,
  stay_id,
  MAX(CASE WHEN (itemid = 50983.0) THEN valuenum END) AS sodium,
  MAX(CASE WHEN (itemid = 50912.0) THEN valuenum END) AS creatinine,
  MAX(CASE WHEN (itemid = 50882.0) THEN valuenum END) AS bicarbonate,
  MAX(CASE WHEN (itemid = 50931.0) THEN valuenum END) AS glucose,
  MAX(CASE WHEN (itemid = 51221.0) THEN valuenum END) AS hematocrit,
  MAX(CASE WHEN (itemid = 50971.0) THEN valuenum END) AS potassium,
  MAX(CASE WHEN (itemid = 51301.0) THEN valuenum END) AS white_blood_cells,
  MAX(CASE WHEN (itemid = 50902.0) THEN valuenum END) AS chloride
FROM (
  SELECT subject_id, itemid, valuenum, stay_id
  FROM (
    SELECT
      q01.*,
      RANK() OVER (PARTITION BY subject_id, stay_id, itemid ORDER BY storetime DESC) AS col0
    FROM (
      SELECT LHS.*, stay_id, intime
      FROM (
        SELECT subject_id, itemid, storetime, valuenum
        FROM arrow_001
        WHERE (itemid IN (50912.0, 50971.0, 50983.0, 50902.0, 50882.0, 51221.0, 51301.0, 509
      ) LHS
      LEFT JOIN dbplyr_aCuTttgfoB
        ON (LHS.subject_id = dbplyr_aCuTttgfoB.subject_id)
    ) q01
    WHERE (storetime < intime)
  ) q01
  WHERE (col01 <= 1)
) q01
GROUP BY subject_id, stay_id
```

```
print(labevents_tble)
```

```
# A tibble: 88,086 x 10
   subject_id stay_id sodium creatinine bicarbonate glucose hematocrit potassium
        <dbl>   <dbl>  <dbl>      <dbl>       <dbl>   <dbl>      <dbl>     <dbl>
 1   10000032  3.96e7    126        0.7          25     102       41.1       6.7
 2   10000690  3.71e7    137        1            26      85       36.1       4.8
 3   10000980  3.98e7    144        2.3          21      89       27.3       3.9
 4   10001217  3.46e7    142        0.5          30      87       37.4       4.1
 5   10001217  3.71e7    142        0.6          22     112       38.1       4.2
 6   10001725  3.12e7    139       NA            NA      NA       NA         4.1
 7   10001843  3.97e7    138        1.3          28     131       31.4       3.9
 8   10001884  3.75e7    130        1.1          30     141       39.7       4.5
 9   10002013  3.91e7    137        0.9          24     288       34.9       3.5
10   10002114  3.47e7    125        3.1          18      95       34.3       6.5
# i 88,076 more rows
# i 2 more variables: white_blood_cells <dbl>, chloride <dbl>
```

Q6. Vitals from charted events

`chartevents.csv.gz` (https://mimic.mit.edu/docs/iv/modules/icu/chartevents/) contains all the charted data available for a patient. During their ICU stay, the primary repository of a patient's information is their electronic chart. The `itemid` variable indicates a single measurement type in the database. The `value` variable is the value measured for `itemid`. The first 10 lines of `chartevents.csv.gz` are

```
zcat < ~/mimic/icu/chartevents.csv.gz | head
```

```
subject_id,hadm_id,stay_id,caregiver_id,charttime,storetime,itemid,value,valuenum,valueuom,wa
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226512,39.4,39.4,kg
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226707,60,60,Inch,0
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226730,152,152,cm,0
10000032,29079034,39553978,18704,2180-07-23 14:00:00,2180-07-23 14:18:00,220048,SR (Sinus Rhy
10000032,29079034,39553978,18704,2180-07-23 14:00:00,2180-07-23 14:18:00,224642,Oral,,,0
10000032,29079034,39553978,18704,2180-07-23 14:00:00,2180-07-23 14:18:00,224650,None,,,0
10000032,29079034,39553978,18704,2180-07-23 14:00:00,2180-07-23 14:20:00,223761,98.7,98.7,°F
10000032,29079034,39553978,18704,2180-07-23 14:11:00,2180-07-23 14:17:00,220179,84,84,mmHg,0
10000032,29079034,39553978,18704,2180-07-23 14:11:00,2180-07-23 14:17:00,220180,48,48,mmHg,0
```

`d_items.csv.gz` (https://mimic.mit.edu/docs/iv/modules/icu/d_items/) is the dictionary for the `itemid` in `chartevents.csv.gz`.

```
zcat < ~/mimic/icu/d_items.csv.gz | head
```

```
itemid,label,abbreviation,linksto,category,unitname,param_type,lownormalvalue,highnormalvalue
220001,Problem List,Problem List,chartevents,General,,Text,,
220003,ICU Admission date,ICU Admission date,datetimeevents,ADT,,Date and time,,
220045,Heart Rate,HR,chartevents,Routine Vital Signs,bpm,Numeric,,
220046,Heart rate Alarm - High,HR Alarm - High,chartevents,Alarms,bpm,Numeric,,
220047,Heart Rate Alarm - Low,HR Alarm - Low,chartevents,Alarms,bpm,Numeric,,
220048,Heart Rhythm,Heart Rhythm,chartevents,Routine Vital Signs,,Text,,
220050,Arterial Blood Pressure systolic,ABPs,chartevents,Routine Vital Signs,mmHg,Numeric,90
220051,Arterial Blood Pressure diastolic,ABPd,chartevents,Routine Vital Signs,mmHg,Numeric,60
220052,Arterial Blood Pressure mean,ABPm,chartevents,Routine Vital Signs,mmHg,Numeric,,
```

We are interested in the vitals for ICU patients: heart rate (220045), systolic non-invasive blood pressure (220179), diastolic non-invasive blood pressure (220180), body temperature in Fahrenheit (223761), and respiratory rate (220210). Retrieve a subset of `chartevents.csv.gz` only containing these items for the patients in `icustays_tble`. Further restrict to the first vital measurement within the ICU stay. The final `chartevents_tble` should have one row per ICU stay and columns for each vital measurement.

```
> chartevents_tble
# A tibble: 94,424 × 7
   subject_id  stay_id heart_rate non_invasive_blood_pressure_systolic non_invasive_blood_pressure_diastolic respiratory_rate temperature_fahrenheit
        <int>    <dbl>      <dbl>                                <dbl>                                 <dbl>           <dbl>                 <dbl>
 1   10000032 39553978         91                                   84                                    48              24                  98.7
 2   10000690 37081114         79                                  107                                    63              23                  97.7
 3   10000980 39765666         77                                  150                                    77              23                  98
 4   10001217 34592300         96                                  167                                    95              11                  97.6
 5   10001217 37067082         86                                  151                                    90              18                  98.5
 6   10001725 31205490         55                                   73                                    56              19                  97.7
 7   10001843 39698942        118                                  112                                    71              17                  97.9
 8   10001884 37510196         38                                  180                                    12              10                  98.1
 9   10002013 39060235         80                                  104                                    70              14                  97.2
10   10002114 34672098        105                                  104                                    81              22                  97.9
# ℹ 94,414 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Hint: Use the Parquet format you generated in Homework 2. For reproducibility, make `chartevents_pq` folder available at the current working directory, for example, by a symbolic link.

```
library(arrow)

chartevents_csv <- "~/mimic/icu/chartevents.csv.gz"
# A folder (not a single file) that will contain Parquet data
chartevents_parquet <- "chartevents_pq"
dataset <- open_dataset(chartevents_csv, format = "csv")
write_dataset(dataset, chartevents_parquet, format = "parquet")
```

```r
itemid_label <- c(
  "220045" = "heart_rate",
  "220179" = "systolic non-invasive blood pressure",
  "220180" = "diastolic non-invasive blood pressure",
  "223761" = "temperature in Fahrenheit",
  "220210" = "respiratory rate"
)
```

```r
chartevents_tble <- open_dataset("chartevents_pq", format = "parquet") |>
  to_duckdb() |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in% c(220045, 220179, 220180, 223761, 220210)) |>
  left_join(
    select(icustays_tble, subject_id, stay_id, intime, outtime),
    by = "subject_id",
    copy = TRUE
  ) |>
  filter(storetime >= intime, storetime <= outtime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_min(storetime, n = 1, with_ties = FALSE) |>
  summarise(valuenum = mean(as.numeric(valuenum), na.rm = TRUE),
            .groups = "drop") |>
  ungroup() |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  rename_at(
    vars(names(itemid_label)),
    ~ itemid_label[.]
  ) |>
  show_query() |>
  collect() |>
  arrange(subject_id, stay_id)
```

```
<SQL>
SELECT
  subject_id,
  stay_id,
  MAX(CASE WHEN (itemid = 223761.0) THEN valuenum END) AS "temperature in Fahrenheit",
  MAX(CASE WHEN (itemid = 220045.0) THEN valuenum END) AS heart_rate,
  MAX(CASE WHEN (itemid = 220179.0) THEN valuenum END) AS "systolic non-invasive blood pressu
  MAX(CASE WHEN (itemid = 220180.0) THEN valuenum END) AS "diastolic non-invasive blood press
  MAX(CASE WHEN (itemid = 220210.0) THEN valuenum END) AS "respiratory rate"
FROM (
```

```
  SELECT subject_id, stay_id, itemid, AVG(CAST(valuenum AS NUMERIC)) AS valuenum
  FROM (
    SELECT subject_id, itemid, storetime, valuenum, stay_id, intime, outtime
    FROM (
      SELECT
        q01.*,
        ROW_NUMBER() OVER (PARTITION BY subject_id, stay_id, itemid ORDER BY storetime) AS co
      FROM (
        SELECT LHS.*, stay_id, intime, outtime
        FROM (
          SELECT subject_id, itemid, storetime, valuenum
          FROM arrow_002
          WHERE (itemid IN (220045.0, 220179.0, 220180.0, 223761.0, 220210.0))
        ) LHS
        LEFT JOIN dbplyr_ljbEizpIEB
          ON (LHS.subject_id = dbplyr_ljbEizpIEB.subject_id)
      ) q01
      WHERE (storetime >= intime) AND (storetime <= outtime)
    ) q01
    WHERE (col01 <= 1)
  ) q01
  GROUP BY subject_id, stay_id, itemid
) q01
GROUP BY subject_id, stay_id
```

```
print(chartevents_tble)
```

```
# A tibble: 94,364 x 7
   subject_id  stay_id temperature in Fahren~1 heart_rate systolic non-invasiv~2
        <dbl>    <dbl>                   <dbl>      <dbl>                   <dbl>
 1   10000032 39553978                    98.7         91                      84
 2   10000690 37081114                    97.7         80                     105
 3   10000980 39765666                    98           75                     158
 4   10001217 34592300                    97.6         69                     145
 5   10001217 37067082                    98.5         86                     151
 6   10001725 31205490                    97.7         86                      73
 7   10001843 39698942                    97.9        131                     108
 8   10001884 37510196                    98.1         60                     167
 9   10002013 39060235                    97.2         80                      93
10   10002114 34672098                    97.9        110                     112
# i 94,354 more rows
# i abbreviated names: 1: `temperature in Fahrenheit`,
```

```
#   2: `systolic non-invasive blood pressure`
# i 2 more variables: `diastolic non-invasive blood pressure` <dbl>,
#   `respiratory rate` <dbl>
```

Q7. Putting things together

Let us create a tibble `mimic_icu_cohort` for all ICU stays, where rows are all ICU stays of adults (age at `intime` >= 18) and columns contain at least following variables

- all variables in `icustays_tble`
- all variables in `admissions_tble`
- all variables in `patients_tble`
- the last lab measurements before the ICU stay in `labevents_tble`
- the first vital measurements during the ICU stay in `chartevents_tble`

The final `mimic_icu_cohort` should have one row per ICU stay and columns for each variable.

```
> mimic_icu_cohort
# A tibble: 94,458 × 41
   subject_id  hadm_id  stay_id first_careunit        last_careunit intime              outtime              los admittime           dischtime           deathtime
        <dbl>    <dbl>    <dbl> <chr>                 <chr>         <dttm>              <dttm>             <dbl> <dttm>              <dttm>              <dttm>
 1   10000032 29079034 39553978 Medical Intensive Car… Medical Inte… 2180-07-23 14:00:00 2180-07-23 23:50:47 0.410 2180-07-23 12:35:00 2180-07-25 17:55:00 NA
 2   10000690 25860671 37081114 Medical Intensive Car… Medical Inte… 2150-11-02 19:37:00 2150-11-06 17:03:17 3.89  2150-11-02 18:02:00 2150-11-12 13:45:00 NA
 3   10000980 26913865 39765666 Medical Intensive Car… Medical Inte… 2189-06-27 08:42:00 2189-06-27 20:38:27 0.498 2189-06-27 07:38:00 2189-07-03 03:00:00 NA
 4   10001217 24597018 37067082 Surgical Intensive Ca… Surgical Int… 2157-11-20 19:18:02 2157-11-21 22:08:00 1.12  2157-11-18 22:56:00 2157-11-25 18:00:00 NA
 5   10001217 27703517 34592300 Surgical Intensive Ca… Surgical Int… 2157-12-19 15:42:24 2157-12-20 14:27:41 0.948 2157-12-18 16:58:00 2157-12-24 14:55:00 NA
 6   10001725 25563031 31205490 Medical/Surgical Inte… Medical/Surg… 2110-04-11 15:52:22 2110-04-12 23:59:56 1.34  2110-04-11 15:08:00 2110-04-14 15:00:00 NA
 7   10001843 26133978 39698942 Medical/Surgical Inte… Medical/Surg… 2134-12-05 18:50:03 2134-12-06 14:38:26 0.825 2134-12-05 00:10:00 2134-12-06 12:54:00 2134-12-06 12:54:00
 8   10001884 26184834 37510196 Medical Intensive Car… Medical Inte… 2131-01-11 04:20:05 2131-01-20 08:27:30 9.17  2131-01-07 20:39:00 2131-01-20 05:15:00 2131-01-20 05:15:00
 9   10002013 23581541 39060235 Cardiac Vascular Inte… Cardiac Vasc… 2160-05-18 10:00:53 2160-05-19 17:33:33 1.31  2160-05-18 07:45:00 2160-05-23 13:30:00 NA
10   10002114 27793700 34672098 Coronary Care Unit (C… Coronary Car… 2162-02-17 23:30:00 2162-02-20 21:16:27 2.91  2162-02-17 22:32:00 2162-03-04 15:16:00 NA
# i 94,448 more rows
# i 30 more variables: admission_type <chr>, admit_provider_id <chr>, admission_location <chr>, discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dttm>, edouttime <dttm>, hospital_expire_flag <dbl>, gender <chr>, anchor_age <dbl>, anchor_year <dbl>,
#   anchor_year_group <chr>, dod <date>, bicarbonate <dbl>, chloride <dbl>, creatinine <dbl>, glucose <dbl>, potassium <dbl>, sodium <dbl>, hematocrit <dbl>, wbc <dbl>,
#   heart_rate <dbl>, non_invasive_blood_pressure_systolic <dbl>, non_invasive_blood_pressure_diastolic <dbl>, respiratory_rate <dbl>, temperature_fahrenheit <dbl>,
#   age_intime <dbl>
# i Use `print(n = ...)` to see more rows
```

```r
mimic_icu_cohort <- icustays_tble %>%
  # 1) Join ICU stays with admissions
  left_join(admissions_tble, by = c("subject_id", "hadm_id")) %>%

  # 2) Join with patients
  left_join(patients_tble, by = "subject_id") %>%

  # 3) Filter to adult ICU stays
  filter(anchor_age >= 18) %>%

  # 4) Join the last lab measurements before ICU (one row per ICU stay)
  left_join(labevents_tble, by = c("subject_id", "stay_id")) %>%
```

```
  # 5) Join the first vitals during ICU (also one row per ICU stay)
  left_join(chartevents_tble, by = c("subject_id", "stay_id"))

mimic_icu_cohort
```

```
# A tibble: 94,458 x 43
   subject_id  hadm_id  stay_id first_careunit last_careunit intime
        <dbl>    <dbl>    <dbl> <chr>          <chr>         <dttm>
 1   10000032 29079034 39553978 Medical Inten~ Medical Inte~ 2180-07-23 14:00:00
 2   10000690 25860671 37081114 Medical Inten~ Medical Inte~ 2150-11-02 19:37:00
 3   10000980 26913865 39765666 Medical Inten~ Medical Inte~ 2189-06-27 08:42:00
 4   10001217 24597018 37067082 Surgical Inte~ Surgical Int~ 2157-11-20 19:18:02
 5   10001217 27703517 34592300 Surgical Inte~ Surgical Int~ 2157-12-19 15:42:24
 6   10001725 25563031 31205490 Medical/Surgi~ Medical/Surg~ 2110-04-11 15:52:22
 7   10001843 26133978 39698942 Medical/Surgi~ Medical/Surg~ 2134-12-05 18:50:03
 8   10001884 26184834 37510196 Medical Inten~ Medical Inte~ 2131-01-11 04:20:05
 9   10002013 23581541 39060235 Cardiac Vascu~ Cardiac Vasc~ 2160-05-18 10:00:53
10   10002114 27793700 34672098 Coronary Care~ Coronary Car~ 2162-02-17 23:30:00
# i 94,448 more rows
# i 37 more variables: outtime <dttm>, los <dbl>, admittime <dttm>,
#   dischtime <dttm>, deathtime <dttm>, admission_type <chr>,
#   admit_provider_id <chr>, admission_location <chr>,
#   discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dttm>, edouttime <dttm>,
#   hospital_expire_flag <dbl>, admission_hour <int>, ...
```

Q8. Exploratory data analysis (EDA)

Summarize the following information about the ICU stay cohort `mimic_icu_cohort` using appropriate numerics or graphs:

- Length of ICU stay `los` vs demographic variables (race, insurance, marital_status, gender, age at intime)

```
library(dplyr)
library(ggplot2)

# By Race
mimic_icu_cohort %>%
  group_by(race) %>%
  summarize(
```

```r
    n_stays = n(),
    median_los = median(los, na.rm = TRUE),
    iqr_los = IQR(los, na.rm = TRUE)
  ) %>%
  arrange(desc(median_los)) %>%
  print(n = Inf)
```

```
# A tibble: 33 x 4
   race                                  n_stays median_los iqr_los
   <chr>                                   <int>      <dbl>   <dbl>
 1 UNABLE TO OBTAIN                         1881       2.36    3.73
 2 UNKNOWN                                  8457       2.27    3.83
 3 ASIAN - KOREAN                             73       2.25    2.64
 4 PORTUGUESE                                425       2.14    3.06
 5 HISPANIC/LATINO - DOMINICAN               746       2.13    3.09
 6 HISPANIC/LATINO - CENTRAL AMERICAN         73       2.11    3.37
 7 BLACK/AFRICAN                             431       2.08    3.12
 8 AMERICAN INDIAN/ALASKA NATIVE             198       2.08    3.61
 9 BLACK/CARIBBEAN ISLAND                    621       2.04    2.87
10 OTHER                                    3134       1.98    2.96
11 SOUTH AMERICAN                            104       1.97    2.73
12 HISPANIC/LATINO - HONDURAN                 88       1.95    2.63
13 WHITE                                   58888       1.94    2.65
14 ASIAN                                    1095       1.92    2.70
15 HISPANIC/LATINO - PUERTO RICAN           1214       1.92    2.78
16 WHITE - RUSSIAN                           980       1.91    2.40
17 WHITE - OTHER EUROPEAN                   2310       1.91    2.64
18 ASIAN - ASIAN INDIAN                      248       1.90    2.73
19 BLACK/AFRICAN AMERICAN                   8677       1.90    2.69
20 HISPANIC/LATINO - SALVADORAN              174       1.90    2.41
21 ASIAN - CHINESE                          1062       1.89    2.52
22 MULTIPLE RACE/ETHNICITY                    74       1.88    1.84
23 ASIAN - SOUTH EAST ASIAN                  408       1.86    2.22
24 HISPANIC/LATINO - GUATEMALAN              227       1.84    2.05
25 HISPANIC/LATINO - CUBAN                   100       1.84    2.84
26 NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER 133       1.84    2.75
27 BLACK/CAPE VERDEAN                        656       1.83    2.67
28 PATIENT DECLINED TO ANSWER                515       1.80    2.04
29 HISPANIC/LATINO - COLUMBIAN               102       1.80    2.73
30 WHITE - BRAZILIAN                         221       1.74    1.95
31 WHITE - EASTERN EUROPEAN                  272       1.74    2.57
32 HISPANIC/LATINO - MEXICAN                  88       1.70    2.04
```

```
33 HISPANIC OR LATINO                              783       1.64    1.99
```

```r
# By Insurance
mimic_icu_cohort %>%
  group_by(insurance) %>%
  summarize(
    n_stays = n(),
    median_los = median(los, na.rm = TRUE),
    iqr_los = IQR(los, na.rm = TRUE)
  ) %>%
  arrange(desc(median_los))
```

```
# A tibble: 6 x 4
  insurance n_stays median_los iqr_los
  <chr>       <int>      <dbl>   <dbl>
1 No charge       8       2.60    3.26
2 Medicare    51819       2.03    2.81
3 Medicaid    14240       1.90    2.82
4 Private     24540       1.88    2.64
5 Other        2328       1.86    2.71
6 <NA>         1523       1.65    2.47
```

```r
# By Marital Status
mimic_icu_cohort %>%
  group_by(marital_status) %>%
  summarize(
    n_stays = n(),
    mean_los = mean(los, na.rm = TRUE),
    sd_los = sd(los, na.rm = TRUE)
  )
```

```
# A tibble: 5 x 4
  marital_status n_stays mean_los sd_los
  <chr>            <int>    <dbl>  <dbl>
1 DIVORCED          6932     3.58   5.16
2 MARRIED          41907     3.59   5.44
3 SINGLE           26785     3.59   5.49
4 WIDOWED          11073     3.18   4.22
5 <NA>              7761     4.64   6.40
```

```r
# By Gender
mimic_icu_cohort %>%
  group_by(gender) %>%
  summarize(
    n_stays = n(),
    median_los = median(los, na.rm = TRUE)
  )
```

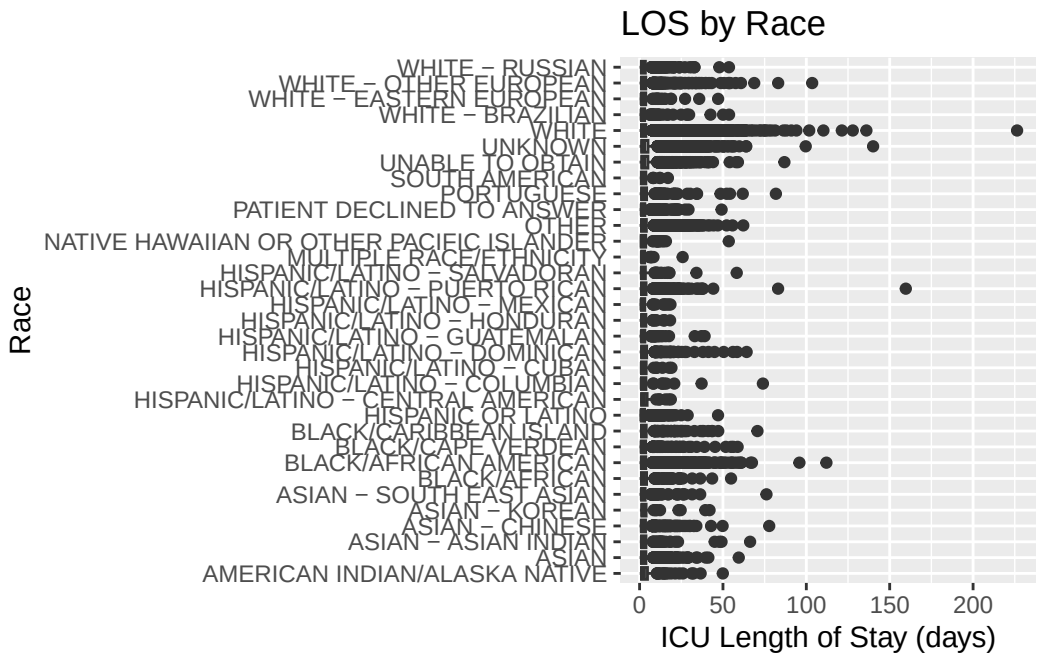
```
# A tibble: 2 x 3
  gender n_stays median_los
  <chr>    <int>      <dbl>
1 F        41583       1.94
2 M        52875       1.98
```

```r
# by age
mimic_icu_cohort %>%
  summarize(
    correlation = cor(anchor_age, los, use = "complete.obs", method = "spearman")
  )
```

```
# A tibble: 1 x 1
  correlation
        <dbl>
1      0.0270
```

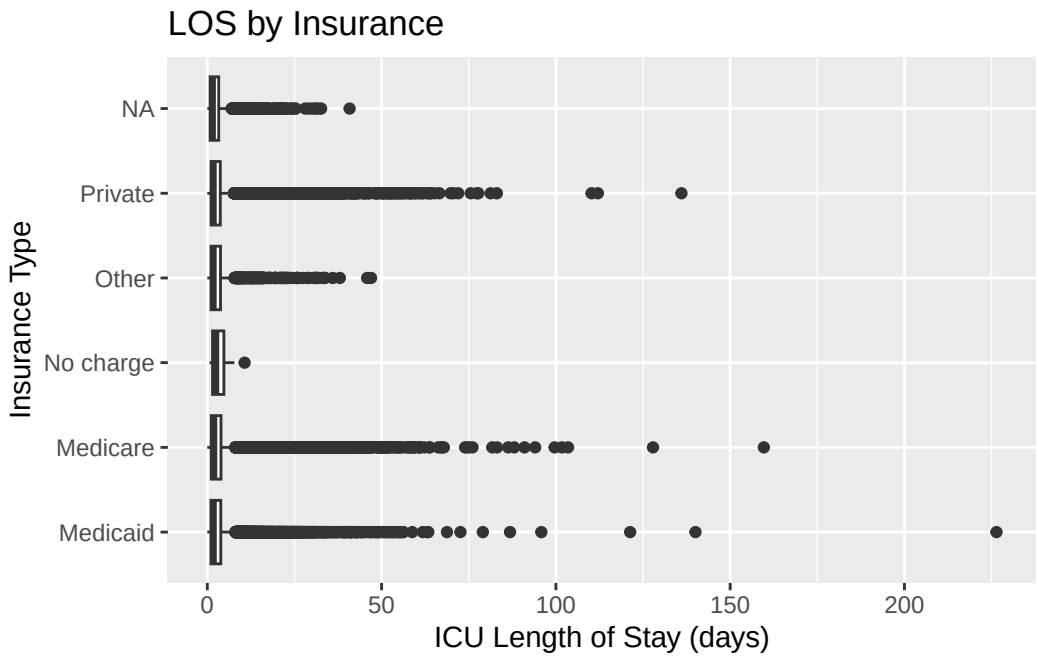
```r
# Boxplots of LOS by categorical demographic variable
ggplot(mimic_icu_cohort, aes(x = race, y = los)) +
  geom_boxplot() +
  coord_flip() +  # Flip for easier reading if many race categories
  labs(title = "LOS by Race",
       x = "Race",
       y = "ICU Length of Stay (days)")
```

```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```




```
ggplot(mimic_icu_cohort, aes(x = insurance, y = los)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "LOS by Insurance",
       x = "Insurance Type",
       y = "ICU Length of Stay (days)")
```

```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```

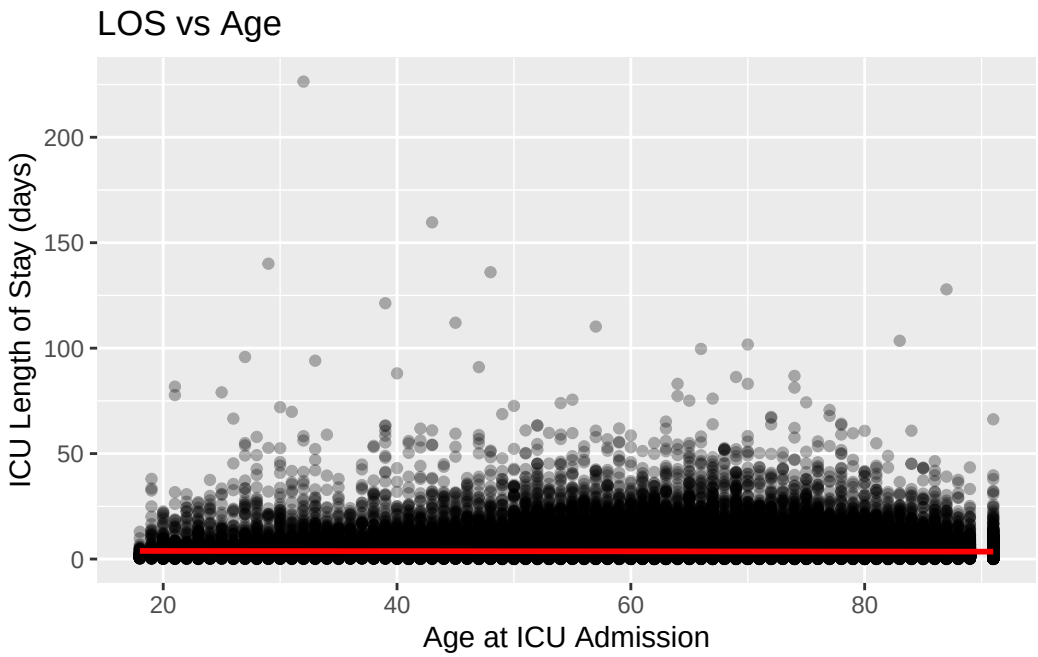



```
# Scatterplot of Age vs LOS
ggplot(mimic_icu_cohort, aes(x = anchor_age, y = los)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "LOS vs Age",
       x = "Age at ICU Admission",
       y = "ICU Length of Stay (days)")
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 14 rows containing missing values or values outside the scale range
(`geom_point()`).
```



- Length of ICU stay `los` vs the last available lab measurements before ICU stay

```
# correlation of each lab vs LOS
labs_of_interest <- c("creatinine", "potassium", "sodium",
                      "chloride", "bicarbonate", "hematocrit",
                      "white_blood_cells", "glucose")

mimic_icu_cohort %>%
  select(los, all_of(labs_of_interest)) %>%
  summarize(across(all_of(labs_of_interest),
                   ~ cor(.x, los, use = "complete.obs", method = "spearman")))
```

```
# A tibble: 1 x 8
  creatinine potassium  sodium chloride bicarbonate hematocrit white_blood_cells
       <dbl>     <dbl>   <dbl>    <dbl>       <dbl>      <dbl>             <dbl>
1     0.0490    0.0121 -0.0214  -0.0402    -0.00972    -0.0545            0.0818
# i 1 more variable: glucose <dbl>
```

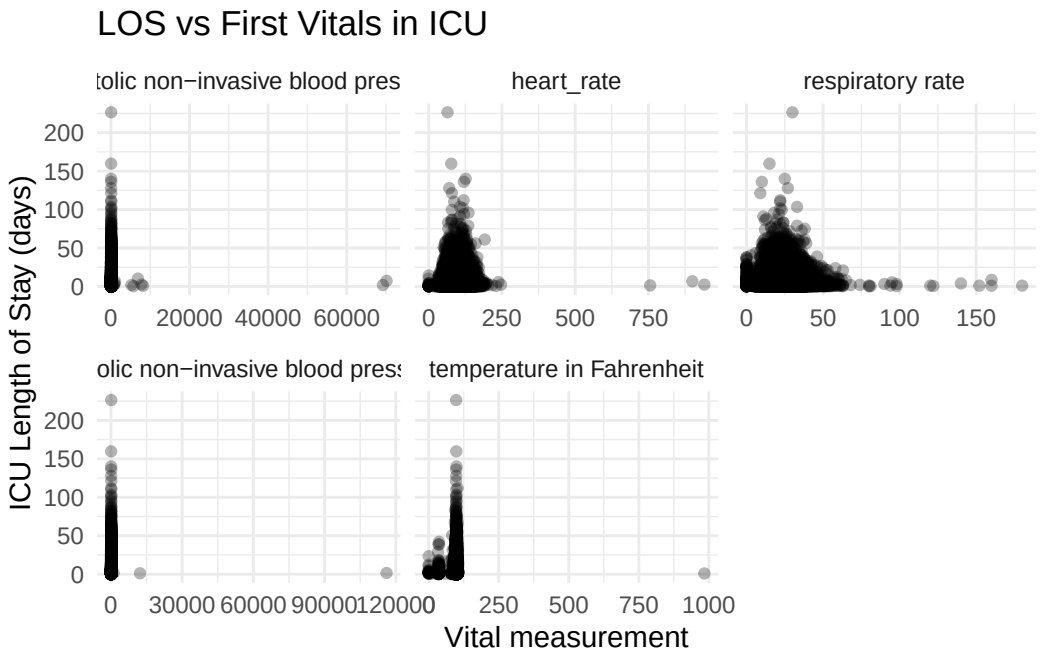
- Length of ICU stay `los` vs the first vital measurements within the ICU stay

```r
vitals_of_interest <- c("heart_rate",
                        "systolic non-invasive blood pressure",
                        "diastolic non-invasive blood pressure",
                        "temperature in Fahrenheit",
                        "respiratory rate")

mimic_icu_cohort %>%
  select(los, all_of(vitals_of_interest)) %>%
  summarize(across(all_of(vitals_of_interest),
                   ~ cor(.x, los, use = "complete.obs", method = "spearman")))
```

```
# A tibble: 1 x 5
  heart_rate `systolic non-invasive blood pressure` diastolic non-invasive blo~1
       <dbl>                                  <dbl>                         <dbl>
1     0.0766                                -0.0561                       -0.0455
# i abbreviated name: 1: `diastolic non-invasive blood pressure`
# i 2 more variables: `temperature in Fahrenheit` <dbl>,
#   `respiratory rate` <dbl>
```

```r
mimic_icu_cohort %>%
  select(los, all_of(vitals_of_interest)) %>%
  pivot_longer(
    cols = all_of(vitals_of_interest),
    names_to = "vital",
    values_to = "value"
  ) %>%
  ggplot(aes(x = value, y = los)) +
  geom_point(alpha = 0.3) +
  facet_wrap(~ vital, scales = "free_x") +  # Each vital on its own panel
  labs(
    title = "LOS vs First Vitals in ICU",
    x = "Vital measurement",
    y = "ICU Length of Stay (days)"
  ) +
  theme_minimal()
```

```
Warning: Removed 4691 rows containing missing values or values outside the scale range
(`geom_point()`).
```



- Length of ICU stay `los` vs first ICU unit

```
mimic_icu_cohort %>%
  group_by(first_careunit) %>%
  summarize(
    n_stays = n(),
    median_los = median(los, na.rm = TRUE),
    iqr_los = IQR(los, na.rm = TRUE)
  ) %>%
  arrange(desc(median_los))
```

```
# A tibble: 17 x 4
   first_careunit                                n_stays median_los iqr_los
   <chr>                                           <int>      <dbl>   <dbl>
 1 Neurology                                           1      28.2     0
 2 Medicine                                           16      13.8     7.92
 3 Surgery/Vascular/Intermediate                     145      13.7    15.1
 4 Surgery/Trauma                                     10      11.6    11.6
 5 Intensive Care Unit (ICU)                          33       5.76   10.2
 6 Neuro Intermediate                               5776       3.00    4.41
 7 Medicine/Cardiology Intermediate                    1       2.58    0
 8 Neuro Surgical Intensive Care Unit (Neuro SICU)  1751       2.24    3.59
```

```
 9 Neuro Stepdown                                    1421       2.20    3.84
10 Coronary Care Unit (CCU)                         10775       2.01    2.68
11 PACU                                               122       2.00    3.21
12 Cardiac Vascular Intensive Care Unit (CVICU)     14771       1.99    2.09
13 Surgical Intensive Care Unit (SICU)              13009       1.98    2.88
14 Medical Intensive Care Unit (MICU)               20703       1.91    2.90
15 Trauma SICU (TSICU)                              10474       1.88    2.77
16 Medical/Surgical Intensive Care Unit (MICU/SICU) 15449       1.79    2.17
17 Med/Surg                                             1       1.44    0
```

```r
ggplot(mimic_icu_cohort, aes(x = first_careunit, y = los)) +
  geom_boxplot() +
  coord_flip() +
  labs(title = "LOS by First ICU Unit",
       x = "First ICU Unit",
       y = "ICU Length of Stay (days)")
```

```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_boxplot()`).
```