Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

AUTHOR
Tanya Wang, 605587605

Display machine information for reproducibility:

```
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0

locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
 [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
[10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

time zone: America/Los_Angeles
tzcode source: system (glibc)

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

loaded via a namespace (and not attached):
 [1] htmlwidgets_1.6.4 compiler_4.4.2    fastmap_1.2.0     cli_3.6.3
 [5] tools_4.4.2       htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10
 [9] rmarkdown_2.29    knitr_1.49        jsonlite_1.8.9    xfun_0.50
[13] digest_0.6.37     rlang_1.1.5       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 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.4     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ purrr::compose()      masks pryr::compose()
✖ lubridate::duration() masks arrow::duration()
✖ tidyr::extract()      masks R.utils::extract()
✖ dplyr::filter()       masks stats::filter()
✖ dplyr::lag()          masks stats::lag()
✖ purrr::partial()      masks pryr::partial()
✖ dplyr::where()        masks pryr::where(), gtsummary::where()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
errors
```

```
library(dplyr)
```

Display your machine memory.

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

```
Totalram:  15.426 GiB
Freeram:   12.791 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.

Solution:

```
ls -l ~/mimic/hosp/labevents.csv.gz
```

```
-rwxrwxrwx 1 tttanyaw tttanyaw 2592909134 Feb  2 21:16 /home/tttanyaw/mimic/hosp/labevents.csv.gz
```

```
labevents <- open_dataset("labevents_pq", format = "parquet")
```

```
id <- 10063848
```

```
labevents <- labevents %>%
  filter(subject_id == id) %>%
  collect()
```

```
path_patients <- "~/mimic/hosp/patients.csv.gz"
path_admissions <- "~/mimic/hosp/admissions.csv.gz"
path_transfers <- "~/mimic/hosp/transfers.csv.gz"
path_procedures <- "~/mimic/hosp/procedures_icd.csv.gz"
path_diagnoses <- "~/mimic/hosp/diagnoses_icd.csv.gz"
path_icd_procedures <- "~/mimic/hosp/d_icd_procedures.csv.gz"
path_icd_diagnoses <- "~/mimic/hosp/d_icd_diagnoses.csv.gz"

patients <- read_csv(path_patients) %>%
  filter(subject_id == id)
```

```
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(path_admissions) %>%
  filter(subject_id == id)
```

```
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(path_transfers) %>%
  filter(subject_id == id)
```

```
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(path_procedures) %>%
  filter(subject_id == id)
```

```
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(path_diagnoses) %>%
  filter(subject_id == id)
```

```
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(path_icd_procedures)
```

```
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(path_icd_diagnoses)
```

```
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.
```

```
# Get top 3 diagnoses
diag <- diagnoses %>%
  left_join(d_icd_diagnoses, by = "icd_code") %>%
  count(long_title, sort = TRUE) %>%
  slice_head(n = 3)
```

```
Warning in left_join(., d_icd_diagnoses, by = "icd_code"): Detected an unexpected many-to-many
relationship between `x` and `y`.
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.
```

```
# Prepare ADT timeline
trans <- transfers %>%
  select(subject_id, careunit, intime, outtime) %>%
  mutate(
    intime = as_date(ymd_hms(intime)),
    outtime = as_date(ymd_hms(outtime)),
    type_of_event = "ADT",
    ICU = ifelse(
      careunit == "Surgical Intensive Care Unit (SICU)", TRUE, FALSE)
  )
```

```
# Prepare lab events
labevent <- labevents %>%
  select(subject_id, charttime) %>%
  mutate(
    charttime = as_date(charttime),
    type_of_event = "Lab") %>%
  rename(event_time = charttime)
```

```
# Prepare procedures
proc <- procedures %>%
  select(subject_id, chartdate, icd_code) %>%
  left_join(d_icd_procedures, by = "icd_code") %>%
  mutate(charttime = as_datetime(chartdate),
         type_of_event = "Procedure") %>%
  rename(event_time = charttime)
```

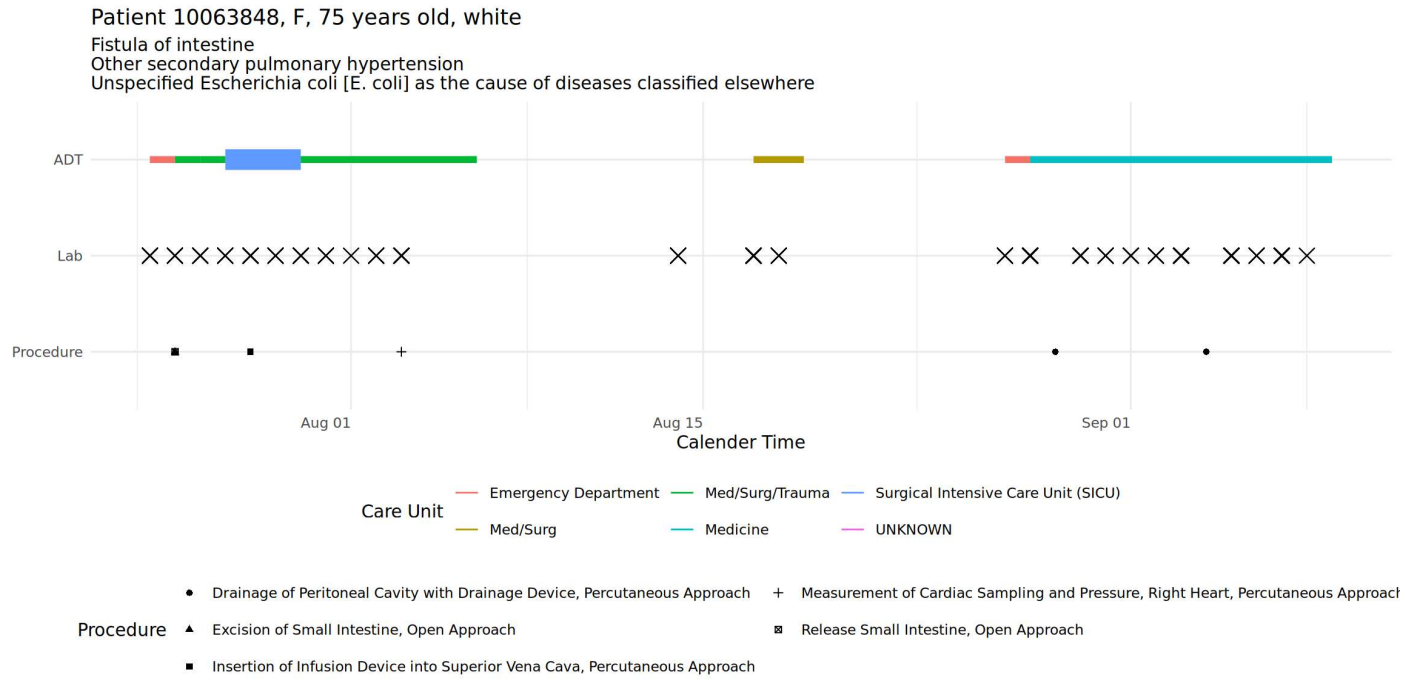
```
# Prepare patient
patient <- patients %>%
  left_join(admissions, by = "subject_id") %>%
  select(subject_id, gender, anchor_age, race) %>%
  unique()
```

```
combine <- bind_rows(
  trans %>%
    select(subject_id, event_time = intime, type_of_event, careunit, ICU),
  trans %>%
    select(subject_id, event_time = outtime, type_of_event, careunit, ICU),
  labevent %>%
    select(subject_id, type_of_event, event_time),
  proc %>%
    select(subject_id, event_time, type_of_event, long_title)) %>%
  mutate (type_of_event =
            factor(type_of_event, levels = c("ADT", "Lab", "Procedure"))
          )

combine <- combine %>%
  mutate(event_time =
           as.POSIXct(event_time, format="%Y-%m-%d %H:%M:%S", tz="UTC"))

trans <- trans %>%
  mutate(
    intime = as.POSIXct(intime, format="%Y-%m-%d %H:%M:%S", tz="UTC"),
    outtime = as.POSIXct(outtime, format="%Y-%m-%d %H:%M:%S", tz="UTC")
  )

labevent <- labevent %>%
  mutate(event_time =
           as.POSIXct(event_time, format="%Y-%m-%d %H:%M:%S", tz="UTC"))
```

```
proc <- proc %>%
  mutate(event_time =
           as.POSIXct(event_time, format="%Y-%m-%d %H:%M:%S", tz="UTC"))
```

```
race_value <- ifelse("race" %in% colnames(patient),
                     tolower(patient$race), "unknown")

plot_title <- paste0("Patient ", patients$subject_id,
                     ", ", patients$gender,
                     ", ", patients$anchor_age,
                     " years old, ", race_value)

plot_subtitle <- paste(na.omit(diag$long_title), collapse = "\n")

ggplot(combine,
       aes(x = event_time, y = type_of_event)) +

  scale_y_discrete(limits = c("Procedure", "Lab", "ADT")) +

  geom_point(
    data = proc,
    aes(x = event_time, y = type_of_event, shape = long_title)) +

  geom_segment(
    data = trans,
    aes(x = intime, xend = outtime, y = type_of_event, yend = type_of_event,
                  color = careunit, linewidth = as.factor(ICU))) +

  geom_point(
    data = labevent,
    aes(x = event_time, y = type_of_event),
    shape = 4, size = 4) +

  labs(
    title = plot_title,
    subtitle = plot_subtitle,
    x = "Calender Time",
    y = NULL,
    color = "Care Unit",
    shape = "Procedure"
  ) +

  guides(
  shape = guide_legend(title = "Procedure", nrow = 3, order = 2),
  color = guide_legend(title = "Care Unit", nrow = 2, order = 1),
  linewidth = "none"
) +

  theme_minimal() +
```

```
  theme(
    legend.position = "bottom",
    legend.box = "vertical",
    axis.text.x = element_text(hjust = 1)
  )
```



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`.

Solution:

```
library(data.table)
```

```
Attaching package: 'data.table'

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

    hour, isoweek, mday, minute, month, quarter, second, wday, week,
    yday, year

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

```
    between, first, last
The following object is masked from 'package:purrr':

    transpose

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

    address
```

```
path_icustays <- "~/mimic/icu/icustays.csv.gz"
path_ditems <- "~/mimic/icu/d_items.csv.gz"

icustays <- fread(path_icustays)
d_items <- fread(path_ditems)
```

```
library(DBI)
library(duckdb)

con <- dbConnect(duckdb::duckdb(), dbdir = ":memory:")

patient_vitals <- dbGetQuery(con, "
    SELECT subject_id, stay_id, charttime, itemid, value
    FROM read_csv_auto('~/mimic/icu/chartevents.csv.gz')
    WHERE subject_id = 10063848
")

dbDisconnect(con)
```

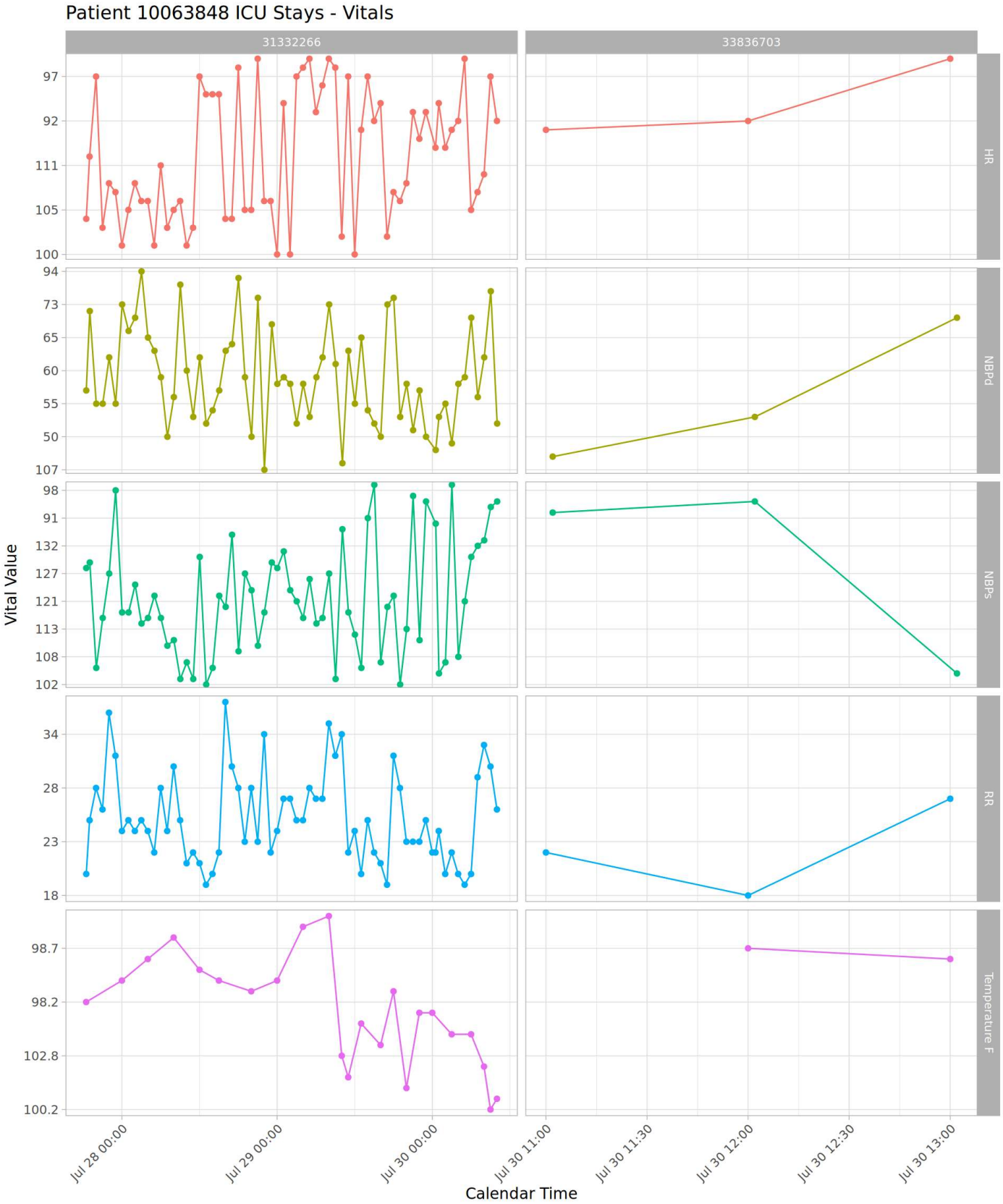
```
id <- 10063848
patient_icustays <- icustays %>%
  filter(subject_id == id) %>%
  select(subject_id, stay_id, intime, outtime)
```

```
vitals <- patient_vitals %>%
  filter(subject_id == id) %>%
  left_join(patient_icustays, by = c("stay_id", "subject_id")) %>%
  select(subject_id, stay_id, charttime, itemid, value) %>%
  mutate(charttime = as.POSIXct(charttime, format="%Y-%m-%d %H:%M:%S"))
```

```
vitals <- vitals %>%
  left_join(d_items %>% select(itemid, abbreviation), by = "itemid") %>%
  rename(vital_type = abbreviation) %>%
  select(subject_id, stay_id, charttime, vital_type, value) %>%
  filter(vital_type %in% c("HR", "RR", "NBPs", "NBPd", "Temperature F"))
```

```
ggplot(vitals, aes(x = charttime, y = value, group = vital_type,
                   color = vital_type)) +
```

```
  geom_line() +

  geom_point(size = 1.5) +

  facet_grid(vital_type ~ stay_id, scales = "free") +

  scale_x_datetime(date_labels = "%b %d %H:%M") +

  scale_y_discrete(breaks = function(x) x[seq(1, length(x), by = 5)]) +

  labs(
    title = paste("Patient", unique(vitals$subject_id), "ICU Stays - Vitals"),
    x = "Calendar Time",
    y = "Vital Value",
    color = "Vital Type"
  ) +

  theme_light() +

  theme(
    strip.text = element_text(size = 8),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```


Patient 10063848 ICU Stays - Vitals
31332266
33836703
HR
NBPd
NBPs
RR
Temperature F
Vital Value
97
92
111
105
100
94
73
65
60
55
50
107
98
91
132
127
121
113
108
102
34
28
23
18
98.7
98.2
102.8
100.2
Jul 28 00:00
Jul 29 00:00
Jul 30 00:00
Jul 30 11:00
Jul 30 11:30
Jul 30 12:00
Jul 30 12:30
Jul 30 13:00
Calendar Time

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
(MICU),2180-07-23 14:00:00,2180-07-23 23:50:47,0.4102662037037037
10000690,25860671,37081114,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit
(MICU),2150-11-02 19:37:00,2150-11-06 17:03:17,3.8932523148148146
10000980,26913865,39765666,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit
(MICU),2189-06-27 08:42:00,2189-06-27 20:38:27,0.4975347222222222
10001217,24597018,37067082,Surgical Intensive Care Unit (SICU),Surgical Intensive Care Unit
(SICU),2157-11-20 19:18:02,2157-11-21 22:08:00,1.1180324074074075
10001217,27703517,34592300,Surgical Intensive Care Unit (SICU),Surgical Intensive Care Unit
(SICU),2157-12-19 15:42:24,2157-12-20 14:27:41,0.948113425925926
10001725,25563031,31205490,Medical/Surgical Intensive Care Unit (MICU/SICU),Medical/Surgical
Intensive Care Unit (MICU/SICU),2110-04-11 15:52:22,2110-04-12 23:59:56,1.338587962962963
10001843,26133978,39698942,Medical/Surgical Intensive Care Unit (MICU/SICU),Medical/Surgical
Intensive Care Unit (MICU/SICU),2134-12-05 18:50:03,2134-12-06 14:38:26,0.8252662037037037
10001884,26184834,37510196,Medical Intensive Care Unit (MICU),Medical Intensive Care Unit
(MICU),2131-01-11 04:20:05,2131-01-20 08:27:30,9.17181712962963
10002013,23581541,39060235,Cardiac Vascular Intensive Care Unit (CVICU),Cardiac Vascular
Intensive Care Unit (CVICU),2160-05-18 10:00:53,2160-05-19 17:33:33,1.314351851851852
```

Q2.1 Ingestion

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

Solution:

```
library(tibble)

icustays_tble <- fread(path_icustays) %>%
  as_tibble()
```

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.

Solution:

Count Unique subject_id:

```
unique_id <- icustays_tble %>%
  summarise(`Unique Subject ID` = n_distinct(subject_id))

print(unique_id)
```

```
# A tibble: 1 × 1
  `Unique Subject ID`
                <int>
1               65366
```

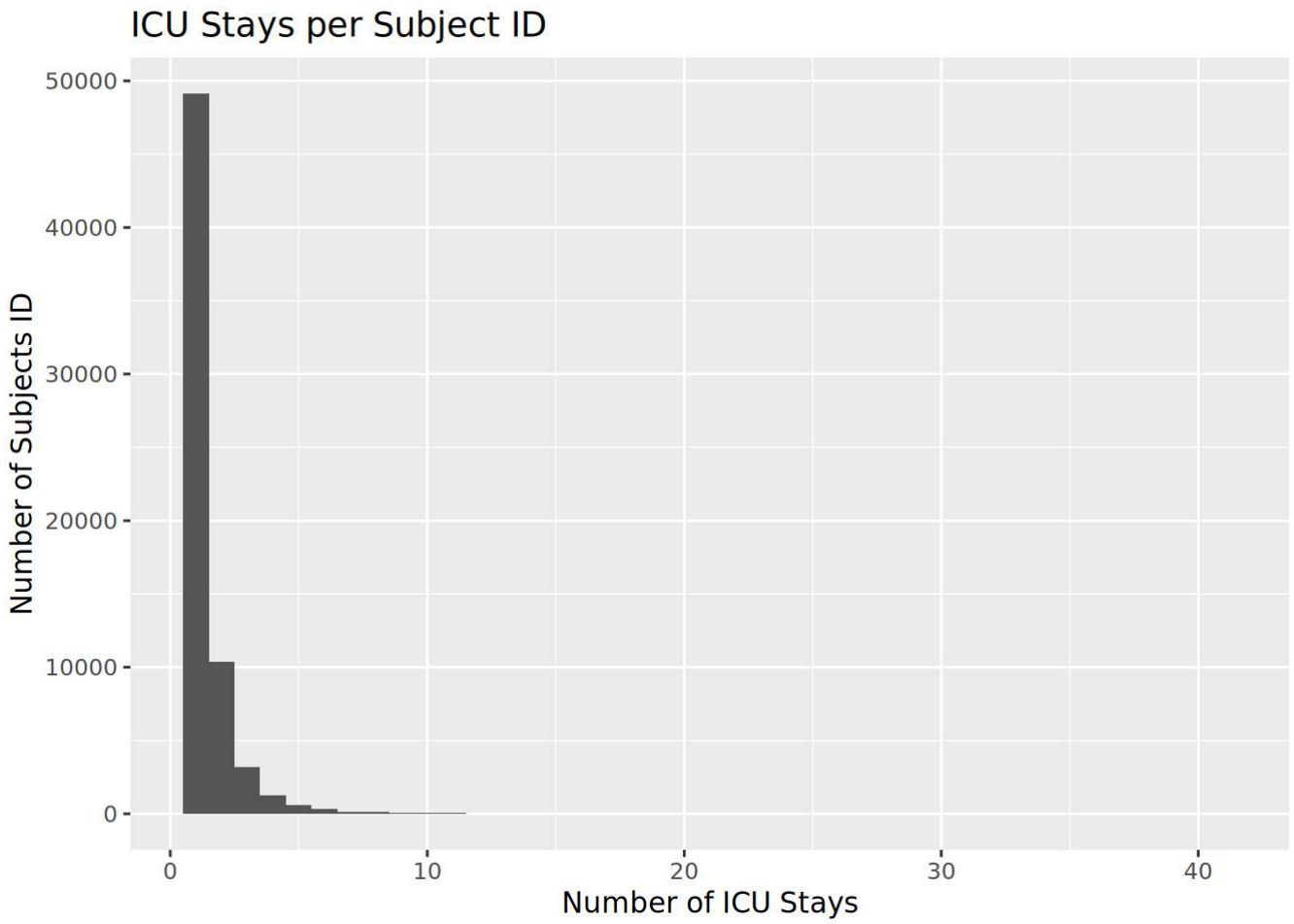
Check if a subject_id can have multiple ICU stays:

```
icu_stay_counts <- icustays_tble %>%
  group_by(subject_id) %>%
  summarise(num_stays = n()) %>%
  ungroup()

icu_stay_counts
```

```
# A tibble: 65,366 × 2
   subject_id num_stays
        <int>     <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
```

Summarize the number of ICU stays per `subject_id` by graphs:

```
ggplot(icu_stay_counts, aes(x = num_stays)) +
  geom_histogram(binwidth = 1) +
  labs(title = "ICU Stays per Subject ID",
       x = "Number of ICU Stays",
       y = "Number of Subjects ID")
```



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_locat
ion,discharge_location,insurance,language,marital_status,race,edregtime,edouttime,hospital_expire
_flag
10000032,22595853,2180-05-06 22:23:00,2180-05-07 17:15:00,,URGENT,P49AFC,TRANSFER FROM
HOSPITAL,HOME,Medicaid,English,WIDOWED,WHITE,2180-05-06 19:17:00,2180-05-06 23:30:00,0
10000032,22841357,2180-06-26 18:27:00,2180-06-27 18:49:00,,EW EMER.,P784FA,EMERGENCY
ROOM,HOME,Medicaid,English,WIDOWED,WHITE,2180-06-26 15:54:00,2180-06-26 21:31:00,0
10000032,25742920,2180-08-05 23:44:00,2180-08-07 17:50:00,,EW EMER.,P19UTS,EMERGENCY
ROOM,HOSPICE,Medicaid,English,WIDOWED,WHITE,2180-08-05 20:58:00,2180-08-06 01:44:00,0
10000032,29079034,2180-07-23 12:35:00,2180-07-25 17:55:00,,EW EMER.,P06OTX,EMERGENCY
ROOM,HOME,Medicaid,English,WIDOWED,WHITE,2180-07-23 05:54:00,2180-07-23 14:00:00,0
10000068,25022803,2160-03-03 23:16:00,2160-03-04 06:26:00,,EU OBSERVATION,P39NWO,EMERGENCY
ROOM,,,English,SINGLE,WHITE,2160-03-03 21:55:00,2160-03-04 06:26:00,0
10000084,23052089,2160-11-21 01:56:00,2160-11-25 14:52:00,,EW EMER.,P42H7G,WALK-IN/SELF
REFERRAL,HOME HEALTH CARE,Medicare,English,MARRIED,WHITE,2160-11-20 20:36:00,2160-11-21
```

```
03:20:00,0
10000084,29888819,2160-12-28 05:11:00,2160-12-28 16:07:00,,EU OBSERVATION,P35NE4,PHYSICIAN
REFERRAL,,Medicare,English,MARRIED,WHITE,2160-12-27 18:32:00,2160-12-28 16:07:00,0
10000108,27250926,2163-09-27 23:17:00,2163-09-28 09:04:00,,EU OBSERVATION,P40JML,EMERGENCY
ROOM,,,English,SINGLE,WHITE,2163-09-27 16:18:00,2163-09-28 09:04:00,0
10000117,22927623,2181-11-15 02:05:00,2181-11-15 14:52:00,,EU OBSERVATION,P47EY8,EMERGENCY
ROOM,,Medicaid,English,DIVORCED,WHITE,2181-11-14 21:51:00,2181-11-15 09:57:00,0
```

Q3.1 Ingestion

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

Solution:

```
admissions_tble <- fread(path_admissions) %>%
  as_tibble()
```

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?)

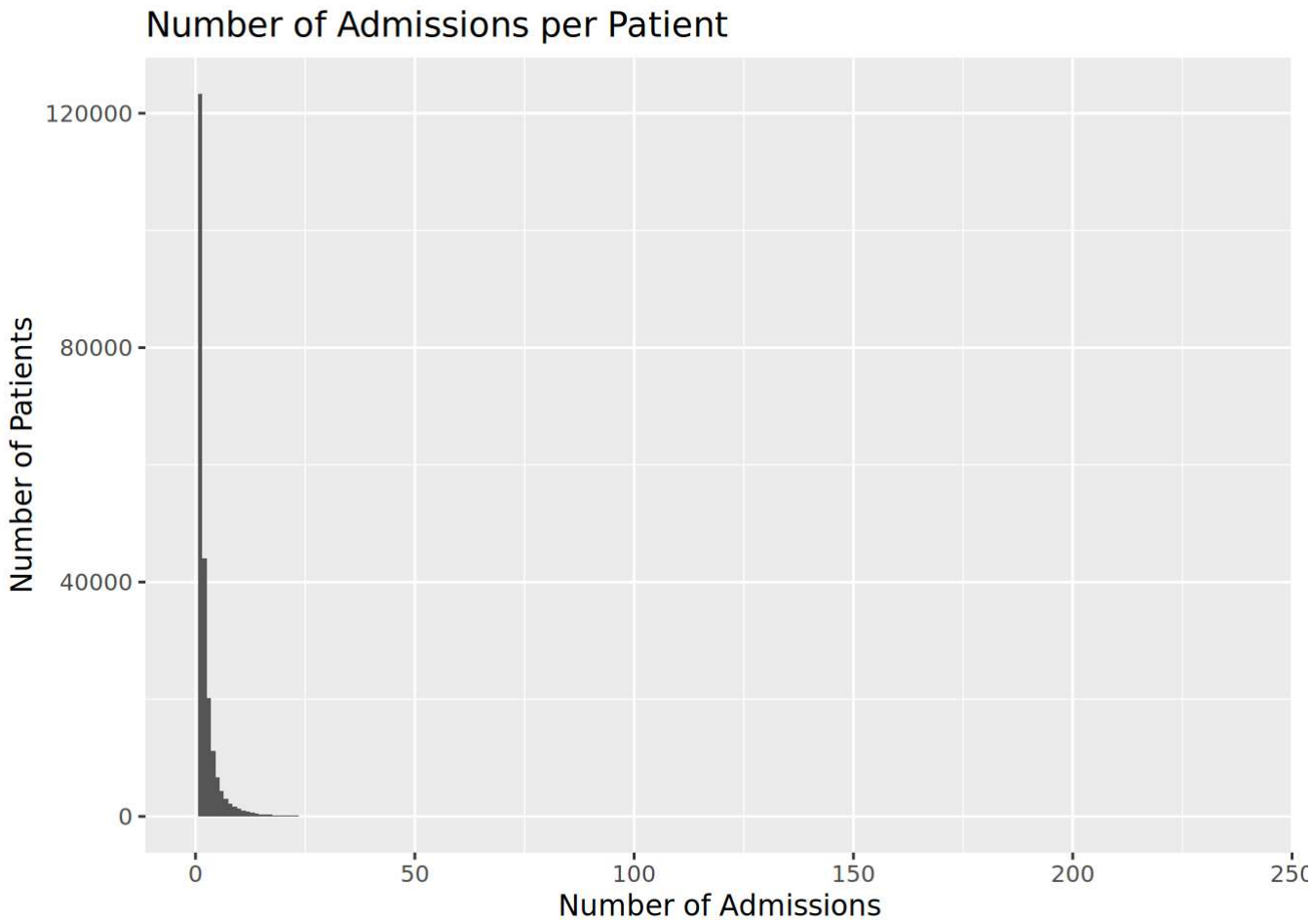
According to the MIMIC-IV documentation,

> 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.

Solution:

number of admissions per patient:

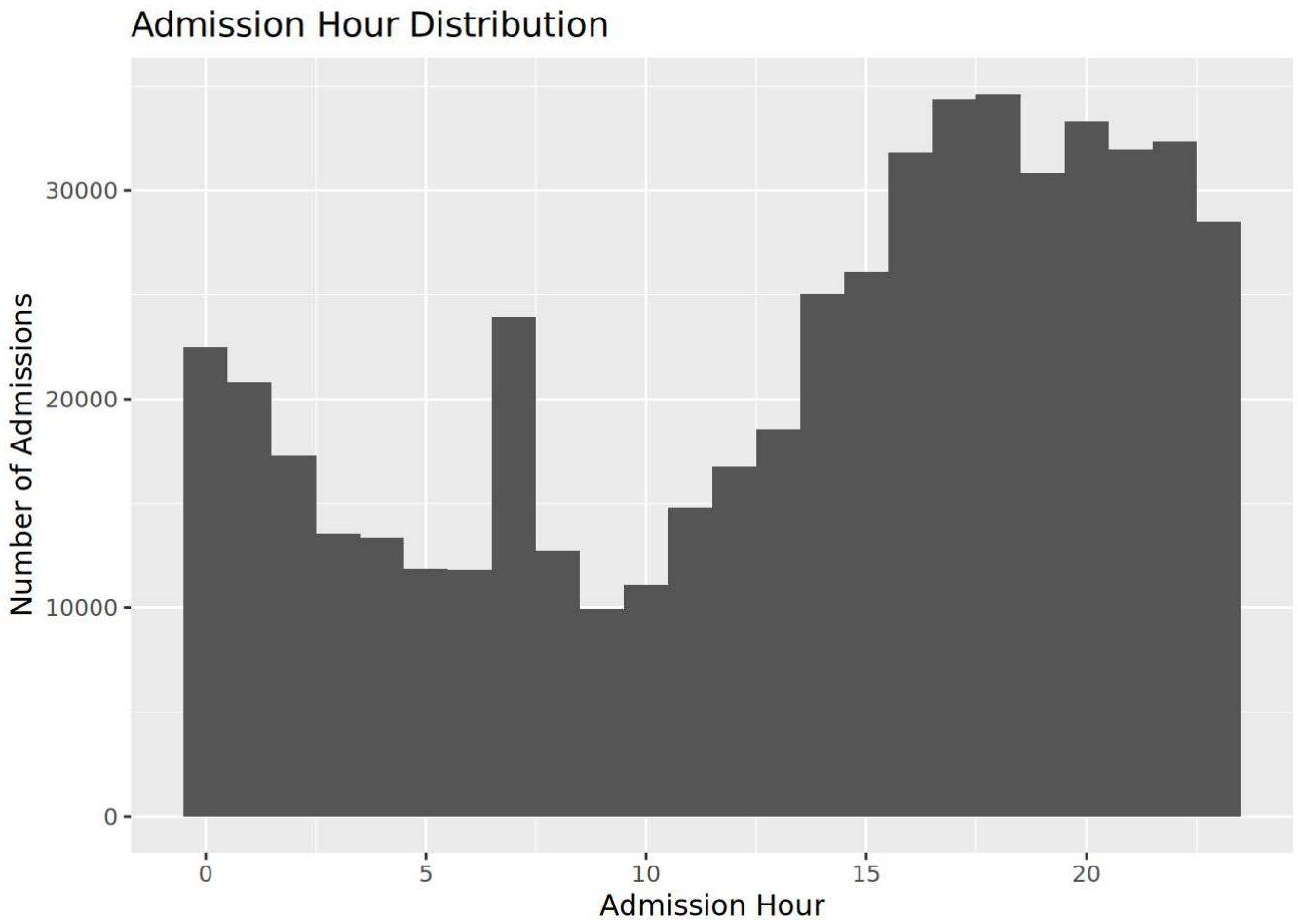
```
admission_count <- admissions_tble %>%
  group_by(subject_id) %>%
  summarise(num_admission = n()) %>%
  ungroup()

ggplot(admission_count, aes(x = num_admission)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Number of Admissions per Patient",
       x = "Number of Admissions",
       y = "Number of Patients")
```



The histogram shows a of right-skewed distribution of hospital admissions per patient. The majority of patients have only one admission, suggesting that most hospital visits are isolated events rather than recurrent ones.

admission hour:

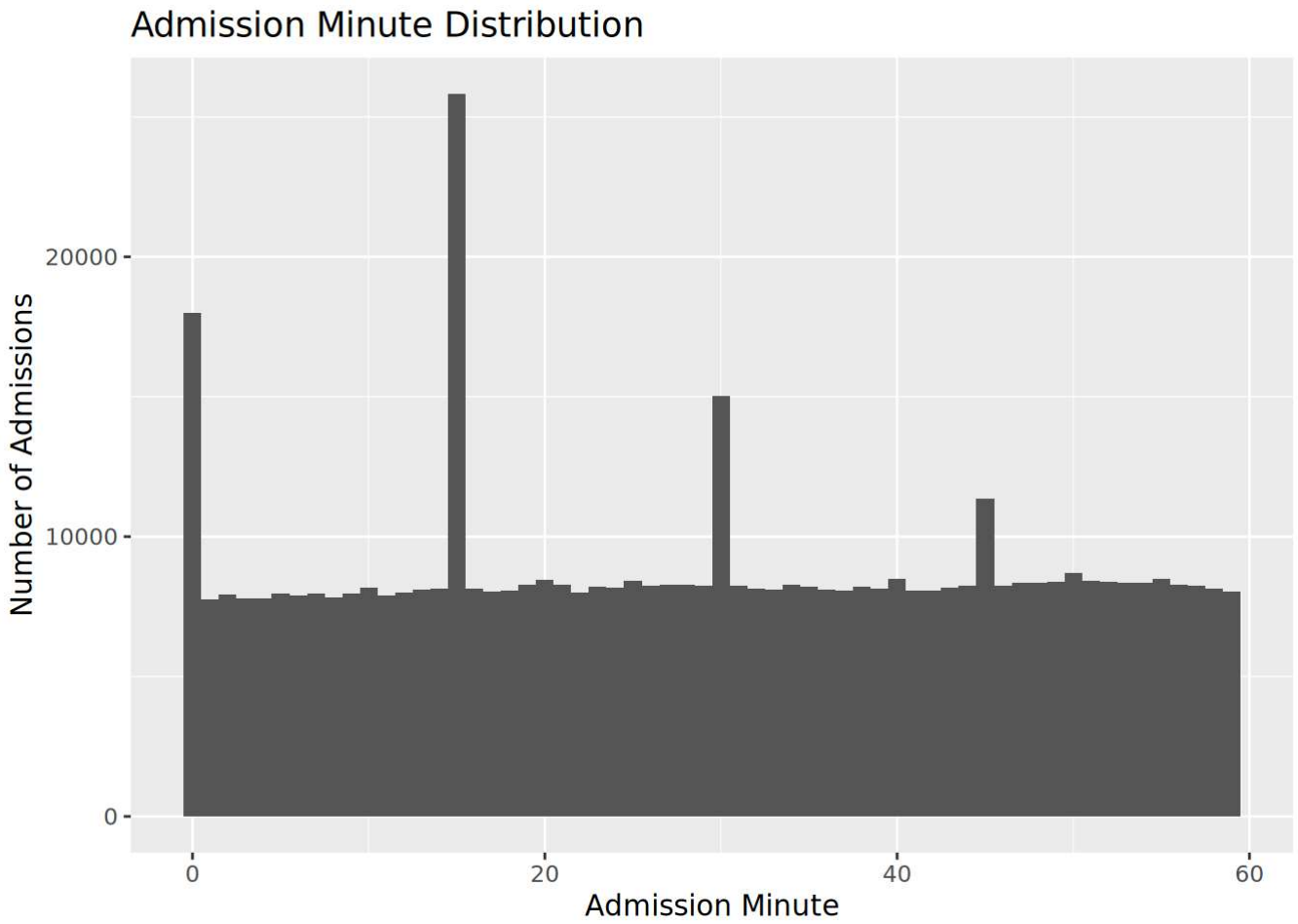
```
admission_hour <- admissions_tble %>%
  mutate(admittime = ymd_hms(admittime)) %>%
  mutate(admit_hour = hour(admittime))

ggplot(admission_hour, aes(x = admit_hour)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Admission Hour Distribution",
       x = "Admission Hour",
       y = "Number of Admissions")
```



According to the plot, the distribution shows a increasing trend of admissions begin at 10 AM, peaking in the evening. Fewer admissions occur overnight (12 AM – 6 AM), except for a spike at 7 AM.

admission minute:

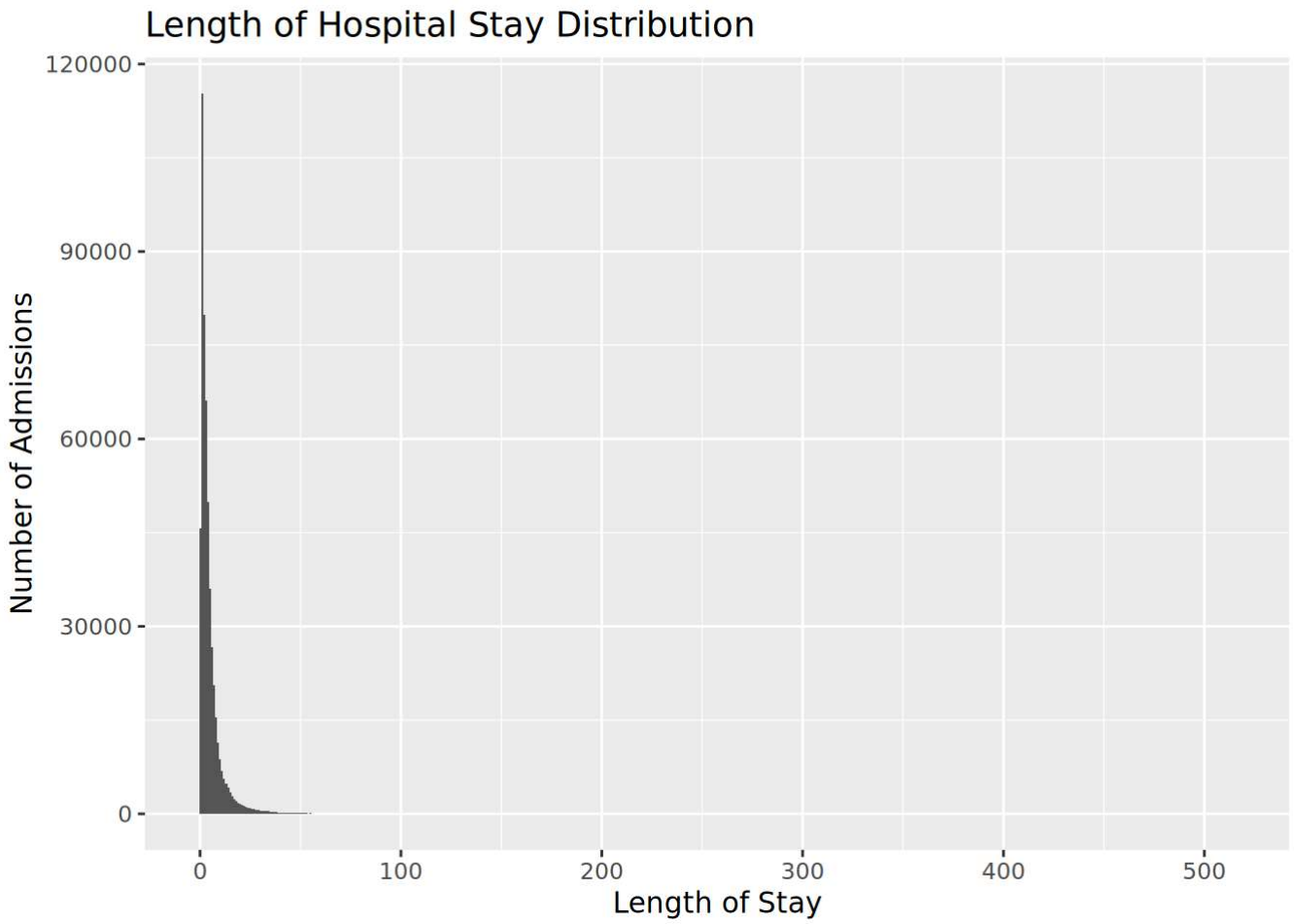
```
admissions_minute <- admissions_tble %>%
  mutate(admittime = ymd_hms(admittime)) %>%
  mutate(admit_minute = minute(admittime))

ggplot(admissions_minute, aes(x = admit_minute)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Admission Minute Distribution",
       x = "Admission Minute",
       y = "Number of Admissions")
```



The admission hour distribution has a structured pattern in hospital admissions. There are several notable spike at whole hour, 15 minutes, 30 minutes, and 45 minutes.

length of hospital stay (from admission to discharge):

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

ggplot(hospital_stay, aes(x = length_stay)) +
  geom_histogram(binwidth = 1) +
  labs(title = "Length of Hospital Stay Distribution",
       x = "Length of Stay",
       y = "Number of Admissions")
```



The length of hospital stay distribution is highly right-skewed, with most patients having short stays and a sharp decline as the length increases. A small number of patients have very long stays (30+ days). There is nothing unusal based on my observation.

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`.

Solution:

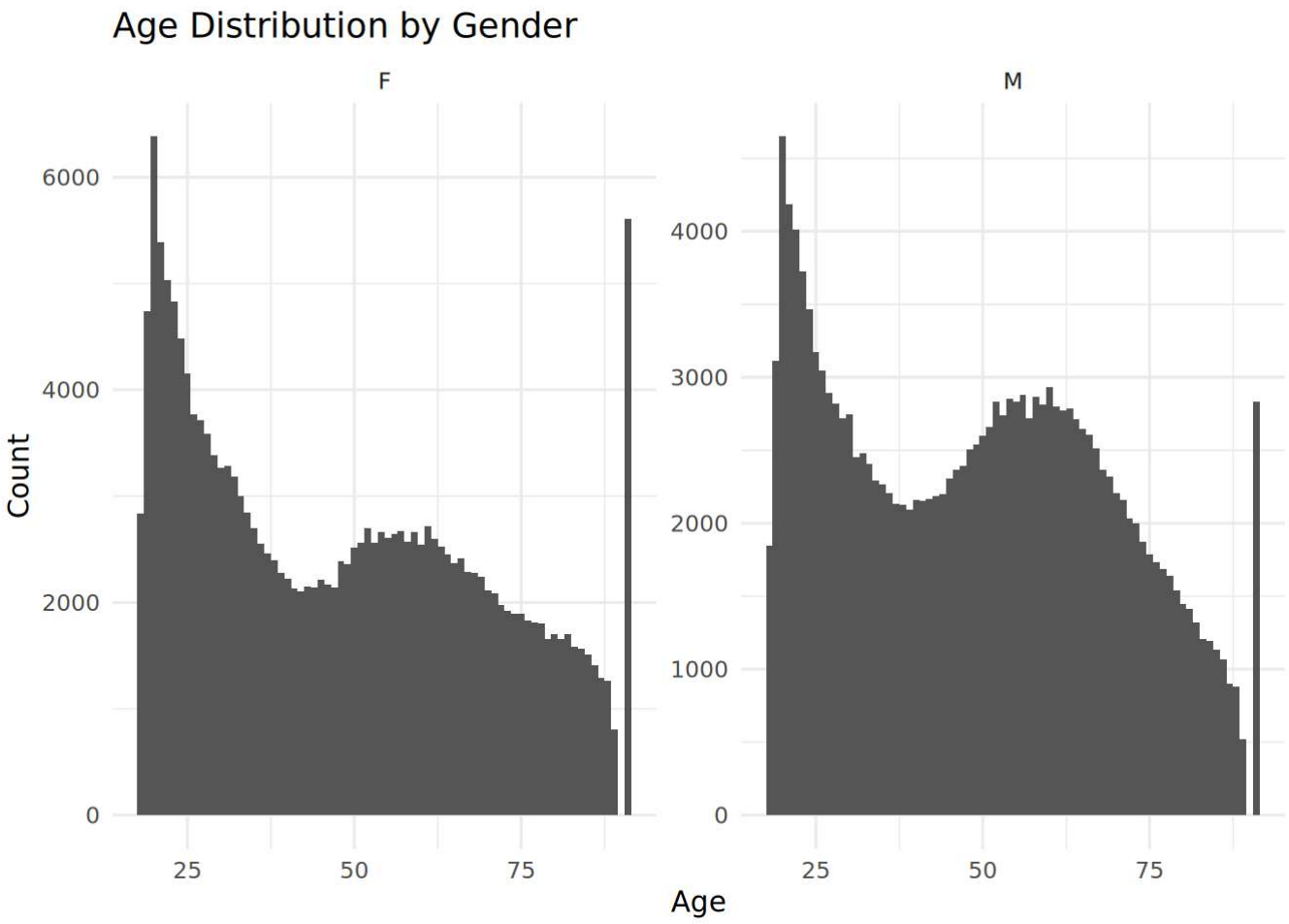
```
patients_tble <- fread(path_patients) %>%
  as_tibble()
```

Q4.2 Summary and visualization

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

Solution:

```
ggplot(patients_tble, aes(x = anchor_age)) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~gender, scales = "free_y") +
  labs(title = "Age Distribution by Gender",
       x = "Age",
       y = "Count") +
  theme_minimal()
```



The plots of female reveals a sharp peak between ages 0-25, followed by a declining trend from 25 to 50 years. Around 50-60 years, there is a slight increase before the trend declines again in older ages. While the overall patterns are similar, males exhibit a broader middle-age peak, suggesting a higher hospitalization rate in this age group. A significant spike at age 90 is observed for both genders, likely due to MIMIC-IV's age-capping policy.

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,value,val
uenum,valueuom,ref_range_lower,ref_range_upper,flag,priority,comments
1,10000032,,2704548,50931,P69FQC,2180-03-23 11:51:00,2180-03-23
15:56:00,___,95,mg/dL,70,100,,ROUTINE,"IF FASTING, 70-100 NORMAL, >125 PROVISIONAL DIABETES."
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,"BENZODIAZEPINE IMMUNOASSAY SCREEN DOES NOT DETECT SOME
DRUGS,;INCLUDING LORAZEPAM, CLONAZEPAM, AND FLUNITRAZEPAM."
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,PRESUMPTIVELY POSITIVE.
8,10000032,,36092842,51090,P69FQC,2180-03-23 11:51:00,2180-03-23
16:00:00,NEG,,,,,,ROUTINE,METHADONE ASSAY DETECTS ONLY METHADONE (NOT OTHER OPIATES/OPIOIDS).
9,10000032,,36092842,51092,P69FQC,2180-03-23 11:51:00,2180-03-23
16:00:00,NEG,,,,,,ROUTINE,"OPIATE IMMUNOASSAY SCREEN DOES NOT DETECT SYNTHETIC OPIOIDS;SUCH AS
METHADONE, OXYCODONE, FENTANYL, BUPRENORPHINE, TRAMADOL,;NALOXONE, MEPERIDINE.  SEE ONLINE LAB
MANUAL FOR DETAILS."
```

`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.

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.

Solution:

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

itemid_values <- as.numeric(names(itemid_map))
```

```
labevents_data  <- open_dataset("labevents_pq", format = "parquet") %>%
  to_duckdb() %>%
  select(subject_id, itemid, storetime, valuenum) %>%
  filter(itemid %in% itemid_values) %>%
  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_with(~ recode(., !!!itemid_map)) %>%
  collect()
```

```
labevents_tble  <- labevents_data %>%
  # Rearrange the columns
  select(
    subject_id,
    stay_id,
```

```
    bicarbonate,
    chloride,
    creatinine,
    glucose,
    potassium,
    sodium,
    hematocrit,
    wbc
  ) %>%
  arrange(subject_id, stay_id) %>%
  as_tibble()

head(labevents_tble, n = 10)
```

```
# A tibble: 10 × 10
   subject_id  stay_id bicarbonate chloride creatinine glucose potassium sodium
        <dbl>    <int>       <dbl>    <dbl>      <dbl>   <dbl>     <dbl>  <dbl>
 1   10000032 39553978          25       95        0.7     102       6.7    126
 2   10000690 37081114          26      100        1        85       4.8    137
 3   10000980 39765666          21      109        2.3      89       3.9    144
 4   10001217 34592300          30      104        0.5      87       4.1    142
 5   10001217 37067082          22      108        0.6     112       4.2    142
 6   10001725 31205490          NA       98       NA        NA       4.1    139
 7   10001843 39698942          28       97        1.3     131       3.9    138
 8   10001884 37510196          30       88        1.1     141       4.5    130
 9   10002013 39060235          24      102        0.9     288       3.5    137
10   10002114 34672098          18       NA        3.1      95       6.5    125
# ℹ 2 more variables: hematocrit <dbl>, wbc <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,warnin
g
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226512,39.4,39.4,kg,0
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
Rhythm),,,0
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,0
```

```
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,90,140
220051,Arterial Blood Pressure diastolic,ABPd,chartevents,Routine Vital Signs,mmHg,Numeric,60,90
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.

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.

Solution:

Decompress `chartevents.csv.gz`:

```
gunzip -c ~/mimic/icu/chartevents.csv.gz > chartevents.csv
echo "chartevents.csv" >> .gitignore
```

Convert `chartevents.csv` to binary Parquet Format:

```
write_dataset(
  open_dataset("chartevents.csv", format = "csv"),
  path = "chartevents_pq",
  format = "parquet"
)
```

```
vitals_map <- c(
  "220045" = "heart_rate",
  "220179" = "non_invasive_blood_pressure_systolic",
  "220180" = "non_invasive_blood_pressure_diastolic",
```

```
  "223761" = "temperature_fahrenheit",
  "220210" = "respiratory_rate"
)

vitals_itemids <- as.numeric(names(vitals_map))
```

```
chartevents_data <- open_dataset("chartevents_pq", format = "parquet") %>%
  to_duckdb() %>%
  select(subject_id, itemid, storetime, valuenum) %>%
  filter(itemid %in% vitals_itemids) %>%
  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, storetime) %>%
  mutate(valuenum = mean(valuenum, na.rm = TRUE)) %>%
  ungroup() %>%
  group_by(subject_id, stay_id, itemid) %>%
  slice_min(order_by = storetime, n = 1) %>%
  select(-storetime, -intime) %>%
  ungroup() %>%
  pivot_wider(names_from = itemid, values_from = valuenum) %>%
  rename_with(~ recode(., !!!vitals_map)) %>%
  collect()
```

```
chartevents_tble  <- chartevents_data %>%
  # Rearrange the columns
  select(
    subject_id,
    stay_id,
    heart_rate,
    non_invasive_blood_pressure_diastolic,
    non_invasive_blood_pressure_systolic,
    respiratory_rate,
    temperature_fahrenheit
  ) %>%
  arrange(subject_id, stay_id) %>%
  as_tibble()

head(chartevents_tble, n = 10)
```

```
# A tibble: 10 × 7
   subject_id  stay_id heart_rate non_invasive_blood_pr…¹ non_invasive_blood_p…²
        <dbl>    <int>      <dbl>                   <dbl>                   <dbl>
 1   10000032 39553978         91                    48                      84
 2   10000690 37081114         78                    56.5                   106
 3   10000980 39765666         76                   102                     154
```

```
 4   10001217 34592300       79.3                    93.3                   156
 5   10001217 37067082       86                      90                     151
 6   10001725 31205490       86                      56                      73
 7   10001843 39698942      124.                     78                     110
 8   10001884 37510196       49                      30.5                   174.
 9   10002013 39060235       80                      62                      98.5
10   10002114 34672098      110.                     80                     112
# ℹ abbreviated names: ¹non_invasive_blood_pressure_diastolic,
#   ²non_invasive_blood_pressure_systolic
# ℹ 2 more variables: respiratory_rate <dbl>, temperature_fahrenheit <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.

Solution:

```
mimic_icu_cohort <- icustays_tble %>%
  left_join(admissions_tble, by = c("subject_id", "hadm_id")) %>%
  left_join(patients_tble, by = "subject_id") %>%
  left_join(labevents_tble, by = c("subject_id", "stay_id")) %>%
  left_join(chartevents_tble, by = c("subject_id", "stay_id")) %>%
  mutate(age_intime = year(intime) - (anchor_year - anchor_age)) %>%
  filter(age_intime >= 18) %>%
  distinct(stay_id, .keep_all = TRUE)

head(mimic_icu_cohort, n = 10)
```

```
# A tibble: 10 × 41
   subject_id  hadm_id  stay_id first_careunit last_careunit intime
        <dbl>    <int>    <int> <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 35 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 <int>, gender <chr>, anchor_age <int>,
#   anchor_year <int>, anchor_year_group <chr>, dod <IDate>, …
```

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)
- Length of ICU stay `los` vs the last available lab measurements before ICU stay
- Length of ICU stay `los` vs the first vital measurements within the ICU stay
- Length of ICU stay `los` vs first ICU unit

Solution:

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

Length of ICU stay `los` vs `race`:

```
mimic_icu_cohort_race <- mimic_icu_cohort %>%
  group_by(race) %>%
  summarise(
    count = n(),
    avg_los = mean(los, na.rm = TRUE),
    median_los = median(los, na.rm = TRUE),
    sd_los = sd(los, na.rm = TRUE)
  ) %>%
  arrange(desc(avg_los))
head(mimic_icu_cohort_race)
```

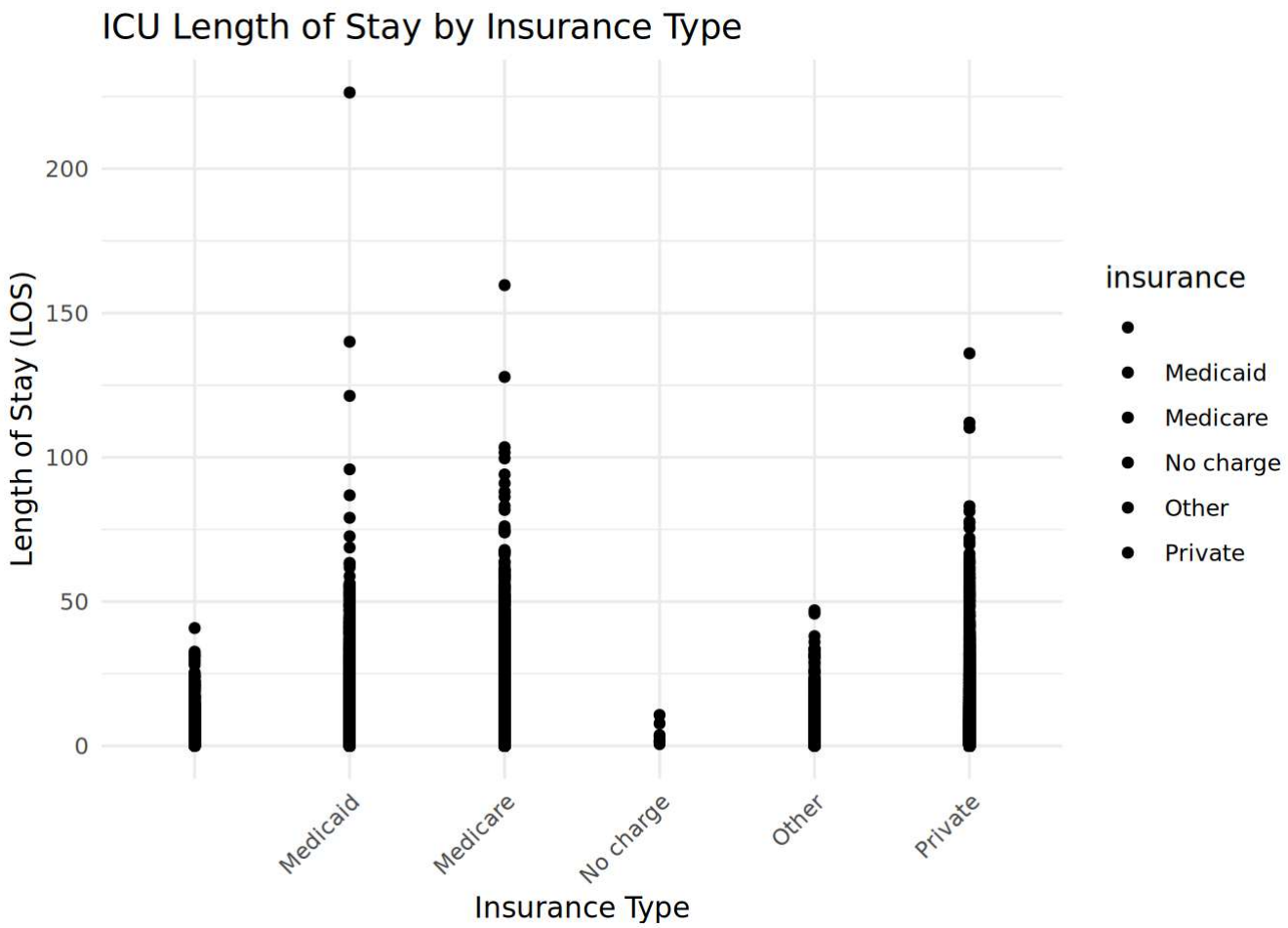
```
# A tibble: 6 × 5
  race                        count avg_los median_los sd_los
  <chr>                       <int>   <dbl>      <dbl>  <dbl>
1 UNABLE TO OBTAIN             1881    4.72       2.36   6.69
2 UNKNOWN                      8457    4.52       2.27   6.33
3 ASIAN - KOREAN                 73    4.44       2.25   7.42
4 PORTUGUESE                    425    4.41       2.14   7.74
```

```
5 BLACK/CARIBBEAN ISLAND          621    4.34       2.04   6.97
6 AMERICAN INDIAN/ALASKA NATIVE   198    4.31       2.08   6.48
```

Length of ICU stay `los` vs `Insurance`:

```
ggplot(mimic_icu_cohort, aes(x = insurance, y = los, fill = insurance)) +
  geom_point() +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "ICU Length of Stay by Insurance Type",
       x = "Insurance Type",
       y = "Length of Stay (LOS)")
```

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

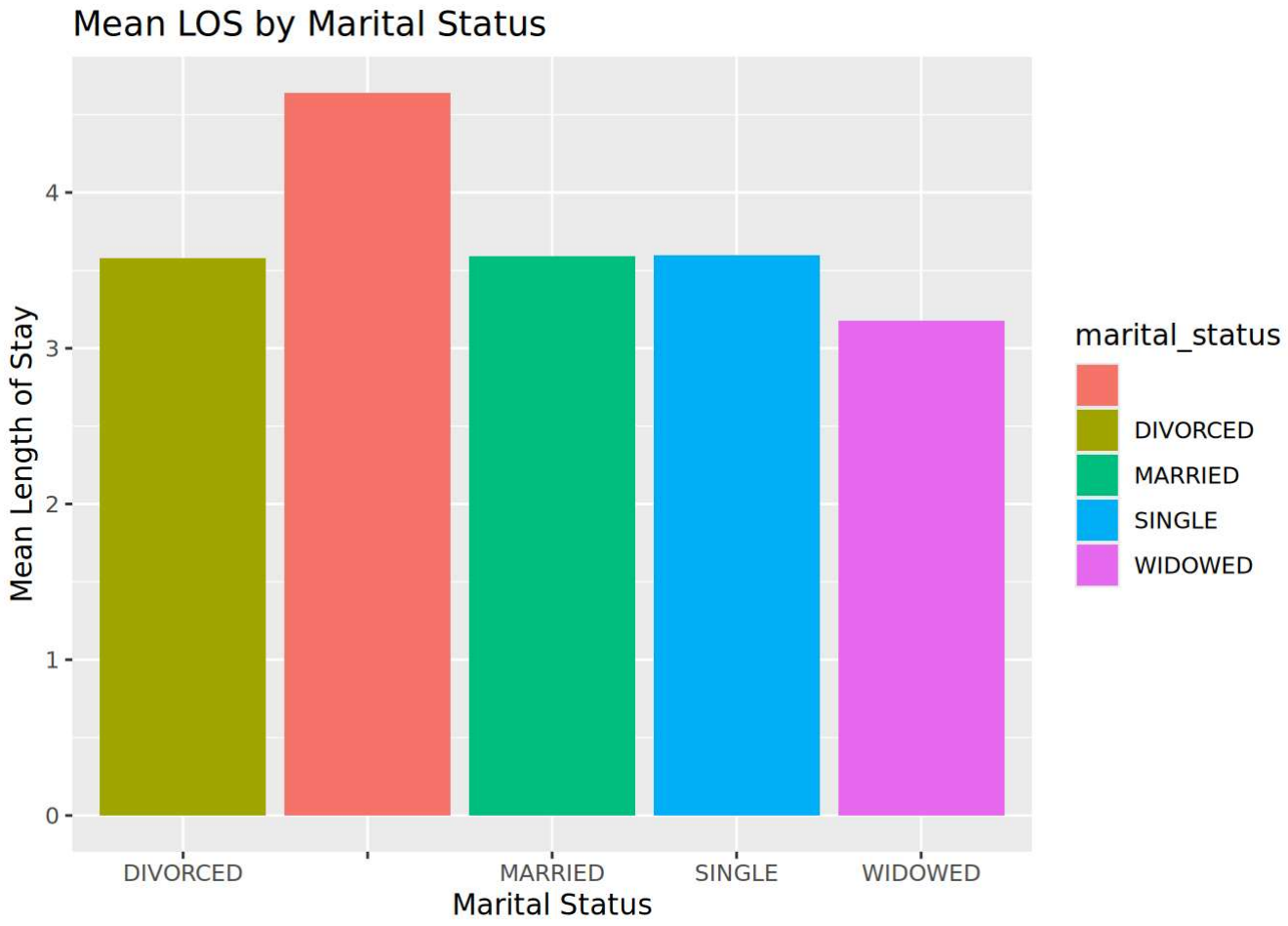


Length of ICU stay `los` vs `marital_status`:

```
ggplot(mimic_icu_cohort,
       aes(x = reorder(marital_status, los, FUN = mean),
           y = los,
           fill = marital_status)) +
  stat_summary(fun = mean, geom = "bar") +
  labs(title = "Mean LOS by Marital Status",
```

```
       x = "Marital Status",
       y = "Mean Length of Stay")
```

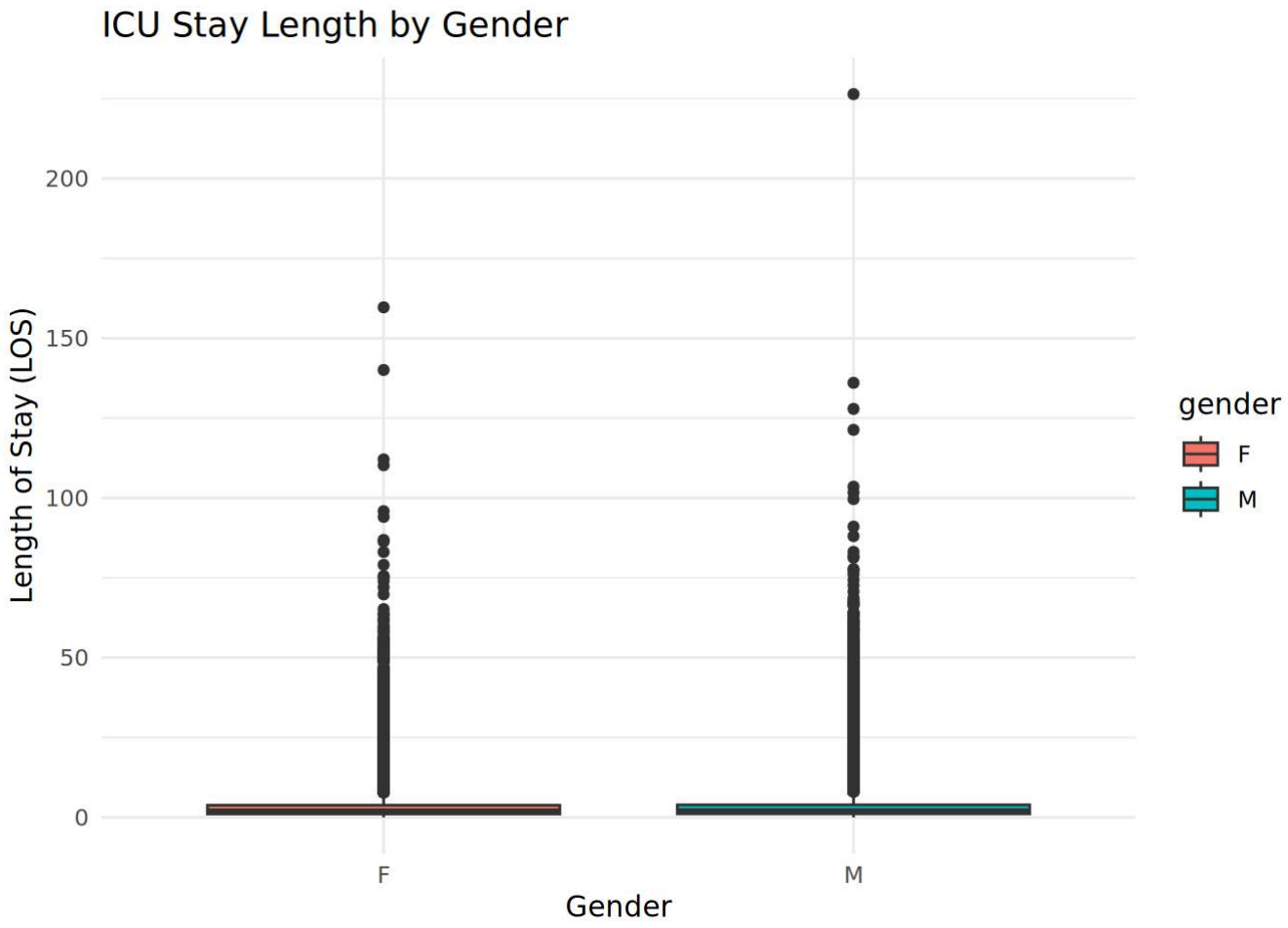
```
Warning: Removed 14 rows containing non-finite outside the scale range
(`stat_summary()`).
```



Length of ICU stay `los` vs `gender`:

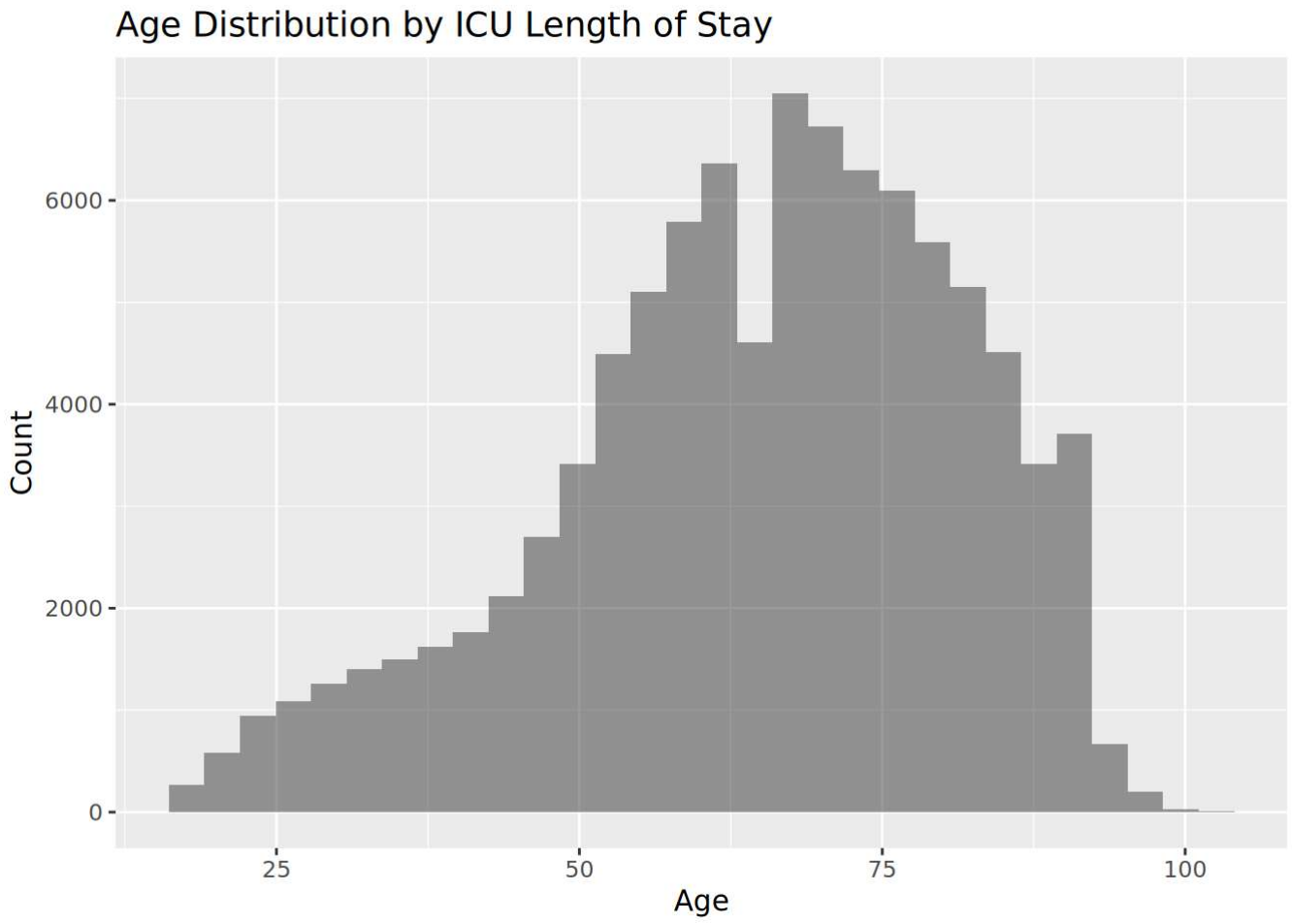
```
ggplot(mimic_icu_cohort, aes(x = gender, y = los, fill = gender)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "ICU Stay Length by Gender",
       x = "Gender",
       y = "Length of Stay (LOS)")
```

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



Length of ICU stay `los` vs `age_intime`:

```
ggplot(mimic_icu_cohort, aes(x = age_intime)) +
  geom_histogram(bins = 30, alpha = 0.6) +
  labs(title = "Age Distribution by ICU Length of Stay",
       x = "Age", y = "Count")
```



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

```
lab_vars <- c("creatinine", "glucose", "wbc", "hematocrit",
              "sodium", "potassium", "chloride", "bicarbonate")

mimic_icu_cohort_2 <- mimic_icu_cohort %>%
  select(los, all_of(lab_vars))
head(mimic_icu_cohort_2)
```

```
# A tibble: 6 × 9
    los creatinine glucose   wbc hematocrit sodium potassium chloride
  <dbl>      <dbl>   <dbl> <dbl>      <dbl>  <dbl>     <dbl>    <dbl>
1 0.410        0.7     102   6.9       41.1    126       6.7       95
2 3.89         1        85   7.1       36.1    137       4.8      100
3 0.498        2.3      89   5.3       27.3    144       3.9      109
4 1.12         0.6     112  15.7       38.1    142       4.2      108
5 0.948        0.5      87   5.4       37.4    142       4.1      104
6 1.34        NA        NA  NA         NA      139       4.1       98
# ℹ 1 more variable: bicarbonate <dbl>
```

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

```
vital_vars <- c("heart_rate",
                "non_invasive_blood_pressure_systolic",
```

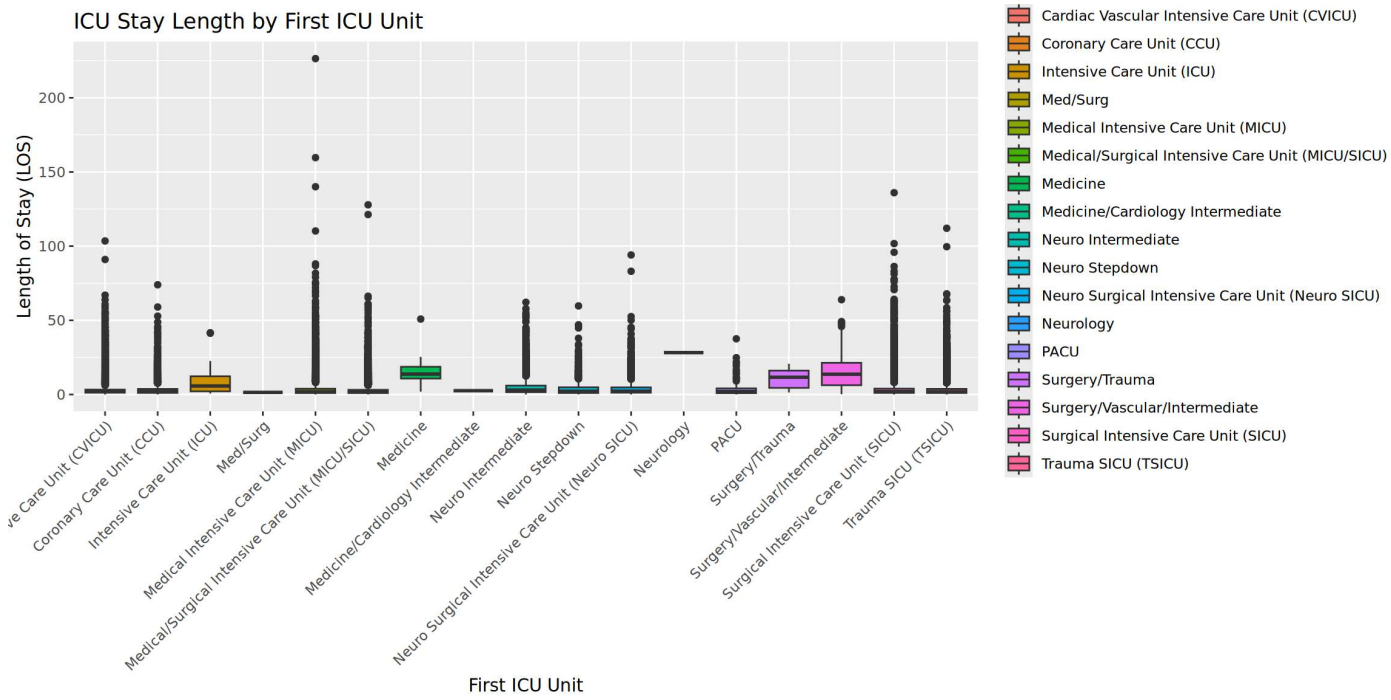
```
                "non_invasive_blood_pressure_diastolic",
                "temperature_fahrenheit",
                "respiratory_rate"
                )

mimic_icu_cohort_3 <- mimic_icu_cohort %>%
  select(los, all_of(vital_vars))
head(mimic_icu_cohort_3)
```

```
# A tibble: 6 × 6
    los heart_rate non_invasive_blood_pressure_systolic non_invasive_blood_pre…¹
  <dbl>      <dbl>                                <dbl>                     <dbl>
1 0.410         91                                   84                      48
2 3.89          78                                  106                      56.5
3 0.498         76                                  154                     102
4 1.12          86                                  151                      90
5 0.948         79.3                                156                      93.3
6 1.34          86                                   73                      56
# ℹ abbreviated name: ¹non_invasive_blood_pressure_diastolic
# ℹ 2 more variables: temperature_fahrenheit <dbl>, respiratory_rate <dbl>
```

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

```
ggplot(mimic_icu_cohort,
       aes(x = first_careunit, y = los, fill = first_careunit)) +
  geom_boxplot() +
  labs(title = "ICU Stay Length by First ICU Unit",
       x = "First ICU Unit",
       y = "Length of Stay (LOS)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

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


ICU Stay Length by First ICU Unit
Length of Stay (LOS)
0
50
100
150
200
ve Care Unit (CVICU)
Coronary Care Unit (CCU)
Intensive Care Unit (ICU)
Med/Surg
Medical Intensive Care Unit (MICU)
Medical/Surgical Intensive Care Unit (MICU/SICU)
Medicine
Medicine/Cardiology Intermediate
Neuro Intermediate
Neuro Stepdown
Neuro Surgical Intensive Care Unit (Neuro SICU)
Neurology
PACU
Surgery/Trauma
Surgery/Vascular/Intermediate
Surgical Intensive Care Unit (SICU)
Trauma SICU (TSICU)
First ICU Unit
Cardiac Vascular Intensive Care Unit (CVICU)
Coronary Care Unit (CCU)
Intensive Care Unit (ICU)
Med/Surg
Medical Intensive Care Unit (MICU)
Medical/Surgical Intensive Care Unit (MICU/SICU)
Medicine
Medicine/Cardiology Intermediate
Neuro Intermediate
Neuro Stepdown
Neuro Surgical Intensive Care Unit (Neuro SICU)
Neurology
PACU
Surgery/Trauma
Surgery/Vascular/Intermediate
Surgical Intensive Care Unit (SICU)
Trauma SICU (TSICU)