Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

AUTHOR
Your Name and UID

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_1.1.4       evaluate_1.0.1
```

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(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
```

```
library(arrow)
library(dplyr)
library(lubridate)
```

```
library(ggplot2)
library(DBI)
library(duckdb)
library(scales)
```

```
Attaching package: 'scales'

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

    discard

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

    col_factor
```

Display your machine memory.

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

```
Totalram:  13.653 GiB
Freeram:   12.552 GiB
```

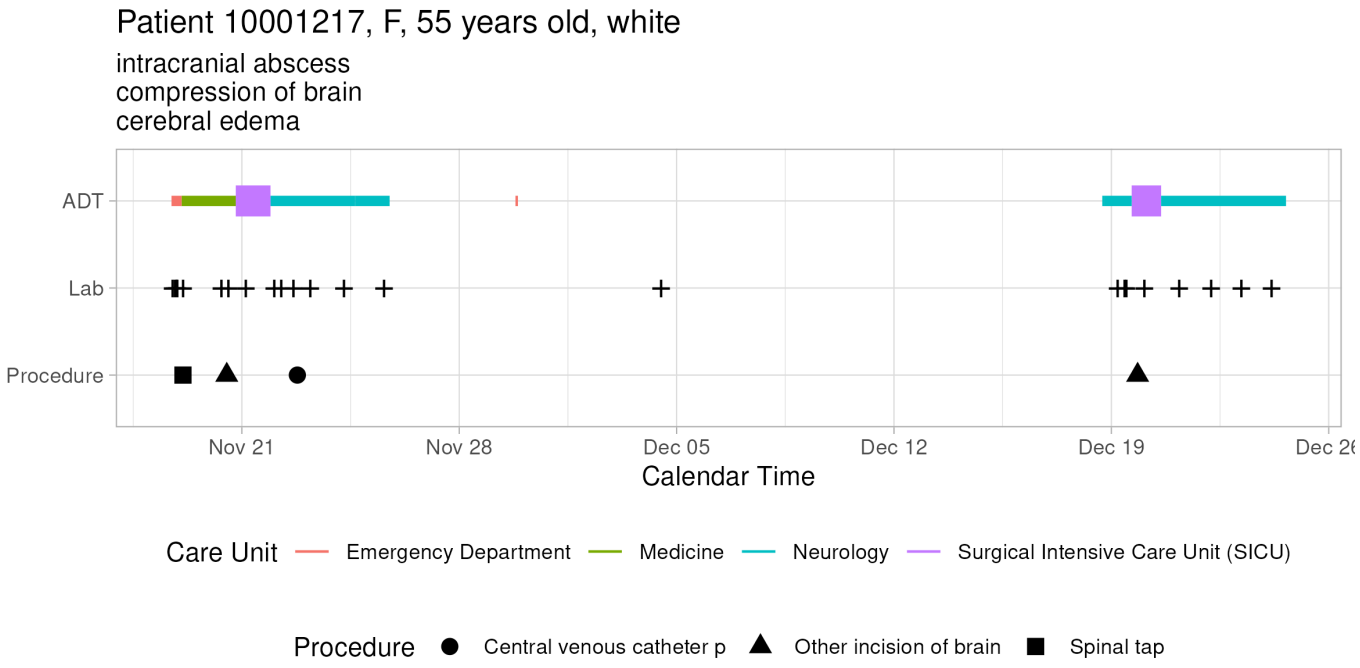
In this exercise, we use tidyverse (ggplot2, dplyr, etc) to explore the MIMIC-IV data introduced in homework 1 and to build a cohort of ICU stays.

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:

Read info with `subject_id = 10063848`:

```
id <- 10063848

# Read info from patients.csv.gz
file_path <- '~/mimic/hosp/patients.csv.gz'
patient <- fread(file_path)
patient <- patient[subject_id == id]

# Read info from admissions.csv.gz
file_path <- '~/mimic/hosp/admissions.csv.gz'
admissions <- fread(file_path)
admissions <- admissions[subject_id == id]

# Read info from transfers.csv.gz
file_path <- '~/mimic/hosp/transfers.csv.gz'
transfers <- fread(file_path)
transfers <- transfers[subject_id == id]

# Read info from labevents.csv.gz
labevents <- open_dataset("labevents_pq", format = "parquet")
labevents <- labevents %>%
```

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

# Read info from procedures_icd.csv.gz
file_path <- '~/mimic/hosp/procedures_icd.csv.gz'
procedures <- fread(file_path)
procedures <- procedures[subject_id == id]

# Read info from diagnoses_icd.csv.gz, d_icd_procedures.csv.gz,
# d_icd_diagnoses.csv.gz
file_paths <- c('~/mimic/hosp/diagnoses_icd.csv.gz',
                '~/mimic/hosp/d_icd_procedures.csv.gz',
                '~/mimic/hosp/d_icd_diagnoses.csv.gz')
diagnoses <- fread(file_paths[1]) %>%
  filter(subject_id == id)

d_icd_procedures <- fread(file_paths[2])
d_icd_diagnoses <- fread(file_paths[3])
```

Merge information needed for the plot:

```
# X-axis: Calender time
# Y-axis: type of event (ADT, lab, procedure)
# Legend: color of line segment (care unit), size (whether ICU or not)
# Legend: crosses of line (lab events)
# Legend: shape of procedure plot (type of procedure)
# Title: patient ID, gender, age, race
# Subtitle: top 3 diagnoses

# Patient information
patient_info <- patient |>
  left_join(admissions, by = "subject_id") |> # get race from admissions
  select(subject_id, gender, anchor_age, race) |>
  distinct()

# Get top 3 diagnoses
top_3_diagnoses <- diagnoses |>
  count(icd_code, sort = TRUE) |>
  arrange(desc(n)) |>
  head(3) |>
  left_join(d_icd_diagnoses, by = "icd_code") |>
  select(long_title)
# top_3_diagnoses_str <- paste(top_3_diagnoses$long_title, collapse = ', ')

# Prepare Lab Events
Lab <- labevents |>
  select(subject_id, charttime) |>
  rename(event_time = charttime) |>
  mutate(event_type = "Lab",
         event_time = as_date(event_time))
```

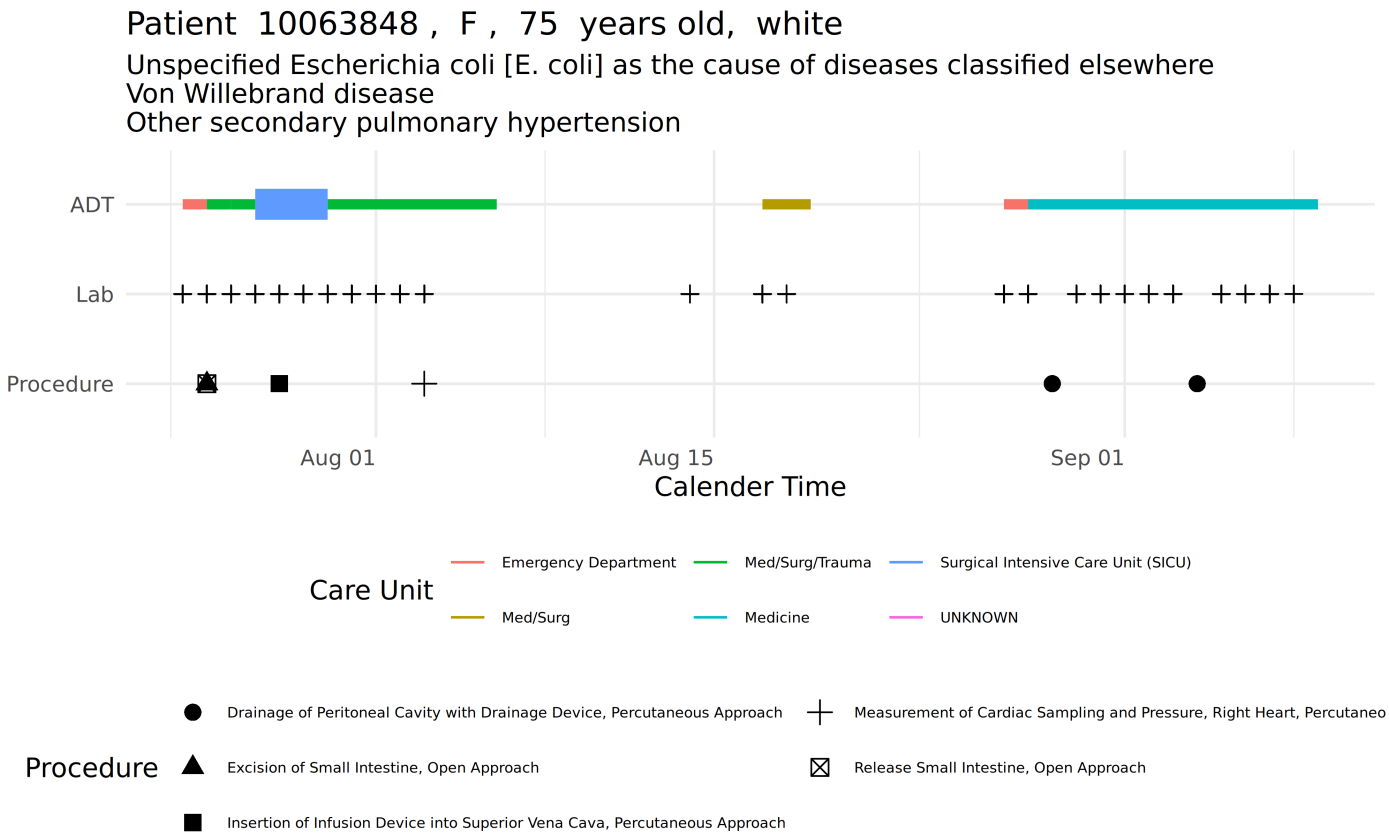
```
# Prepare ADT
ADT <- transfers |>
  select(subject_id, careunit, intime, outtime) |>
  mutate(
    intime = as_date(ymd_hms(intime)),  # Convert to date
    outtime = as_date(ymd_hms(outtime)),  # Convert to date
    event_type = "ADT",
    is_icu = ifelse(careunit == "Surgical Intensive Care Unit (SICU)",
                    TRUE, FALSE)
  )

# Prepare Procedures
proc <- procedures |>
  select(subject_id, chartdate, icd_code) |>
  left_join(d_icd_procedures, by = 'icd_code') |>
  rename(event_time = chartdate,
         procedure_type = long_title) |>
  mutate(event_type = "Procedure",
         event_time = as_date(event_time))

# Combine all events into single dataset
events <- bind_rows(
  ADT |> select(subject_id, event_time = intime, event_type, careunit, is_icu),
  ADT |> select(subject_id, event_time = outtime, event_type, careunit, is_icu),
  Lab |> select(subject_id, event_time, event_type),
  proc |> select(subject_id, event_time, event_type, procedure_type)
) |>
  mutate (event_type = factor(
    event_type, levels = c("ADT", "Lab", "Procedure")
  ))
```

Plot the figure:

```
plot_title <- paste(
  "Patient ", patient_info$subject_id,
  ", ", patient_info$gender,
  ", ", patient_info$anchor_age,
  " years old, ", tolower(patient_info$race)
)
plot_subtitle <- paste(
  top_3_diagnoses$long_title[1], "\n",
  top_3_diagnoses$long_title[2], "\n",
  top_3_diagnoses$long_title[3],
  sep = ""
)
```

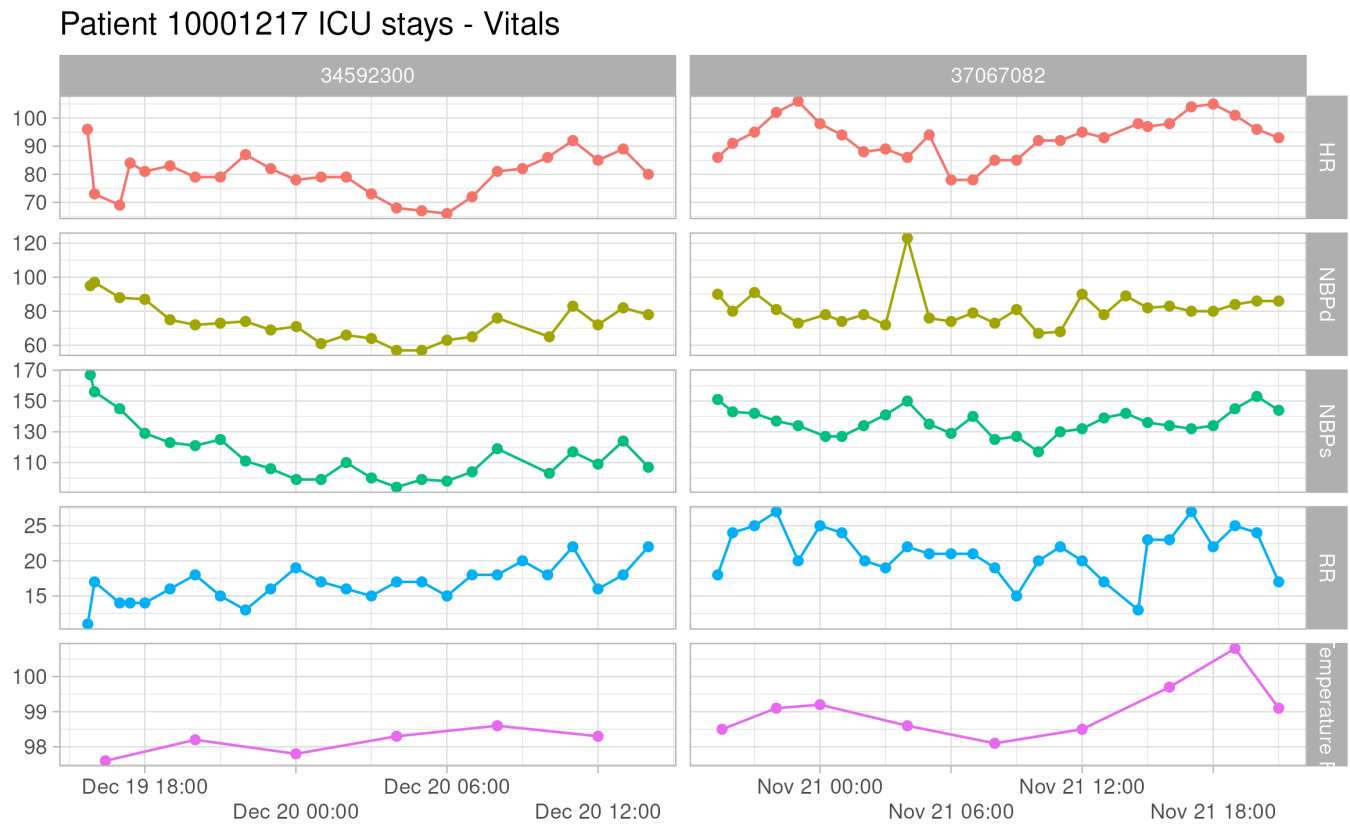
```
ggplot(events,
       aes(x = event_time, y = event_type)) +

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

```
# Procedure Events as different shapes
geom_point(
  data = proc,
  aes(x = event_time, y = event_type, shape = procedure_type),
  size = 3) +

# ADT as line segment
geom_segment(
  data = ADT,
  aes(x = intime, xend = outtime, y = event_type, yend = event_type,
      color = careunit, linewidth = as.factor(is_icu))) +

# Lab Events as Crosses
geom_point(
  data = Lab,
  aes(x = event_time, y = event_type),
  shape = 3, size = 2) +

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

# Legend settings: color and shape
guides(
  # Removes the legend for the line thickness (linewidth)
  linewidth = "none",
  # Title for shape legend
  shape = guide_legend(title = "Procedure", nrow = 3, order = 2),
  # Title for color legend
  color = guide_legend(title = "Care Unit", nrow = 2, order = 1)
) +

theme_minimal() +
theme(
  legend.position = "bottom",
  legend.box = "vertical",  # Stack legends horizontally
  legend.text = element_text(size = 6),
  # legend.spacing.y = unit(0.1, "cm"),
  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:

Read info with `subject_id = 10063848`:

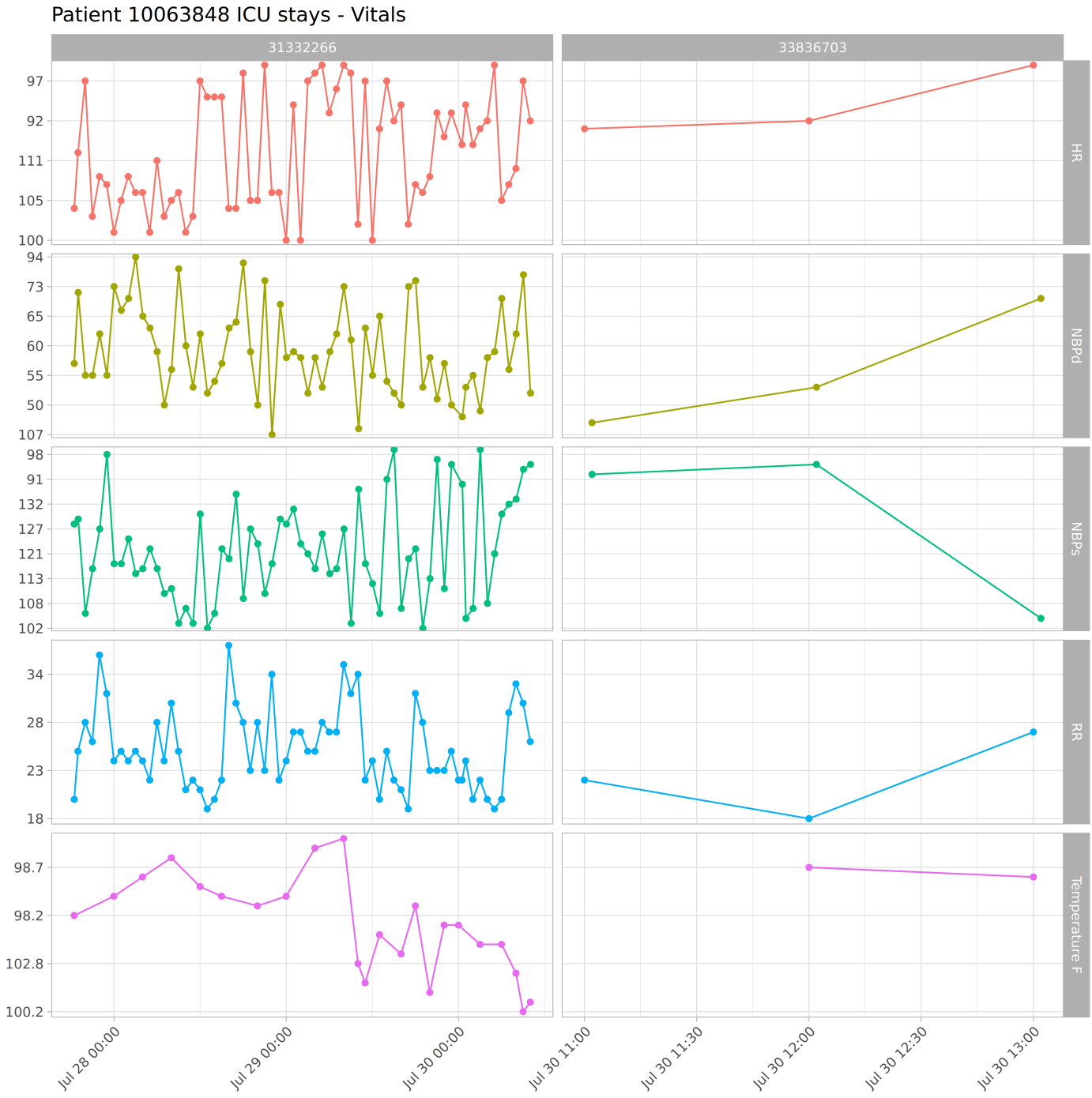
```
id <- 10063848

# Read info from d_items.csv.gz
file_path <- "~/mimic/icu/d_items.csv.gz"
d_items <- fread(file_path)
```

```
# Read info from chartevents.csv.gz
con <- dbConnect(duckdb::duckdb(), dbdir = ":memory:")

chartevents <- 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)
```

Merge information for plotting:

```r
# vitals information
vitals <- chartevents |>
  left_join(d_items, by = "itemid") |>
  filter(abbreviation %in% c("HR", "NBPd", "NBPs", "RR", "Temperature F")) |>
  select(subject_id, stay_id, charttime, value, abbreviation) |>
  mutate(charttime = as.POSIXct(charttime, format = "%Y-%m-%d %H:%M:%S"))
```

```r
# x-axis: calender time
# y-axis: value of vitals
# color: type of vitals
# facet grid: stay ID

ggplot(vitals, aes(x = charttime, y = value, group = abbreviation,
                   color = abbreviation)) +
  geom_point() +
  geom_line() +
  facet_grid(abbreviation ~ stay_id, scales = "free") +

  labs(
    title = paste("Patient", unique(vitals$subject_id),
                  "ICU stays - Vitals", sep = " ")
  ) +

  theme_light() +
  theme(
    legend.position = "none",
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +

  scale_x_datetime(date_labels = "%b %d %H:%M") +
  scale_y_discrete(breaks = function(x) x[seq(1, length(x), by = 5)])
```



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:

```
# Read info from icustays.csv.gz
file_path <- "~/mimic/icu/icustays.csv.gz"
icustays_tble <- fread(file_path) |> 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:

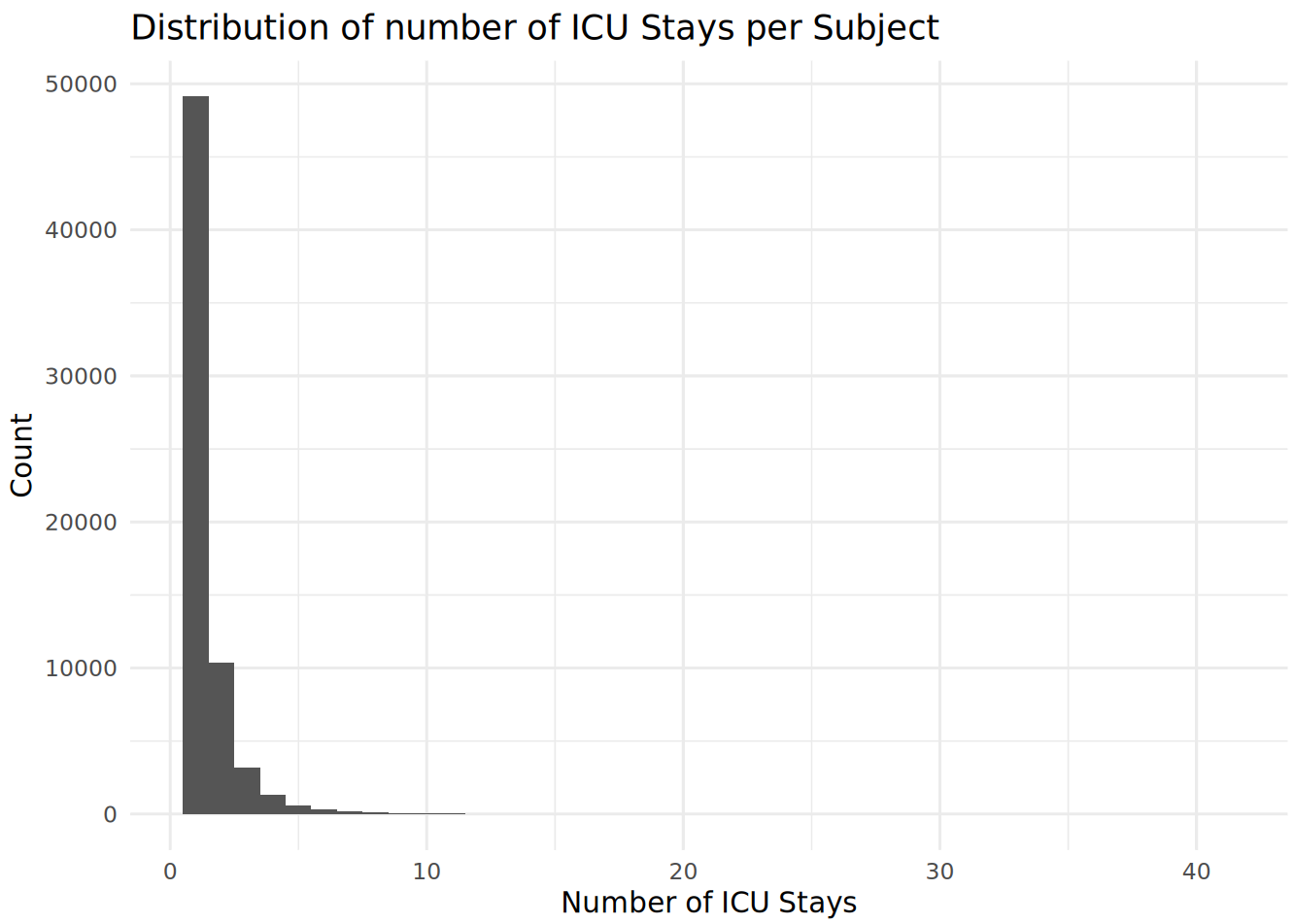
```
length(unique(icustays_tble$subject_id))
```

```
[1] 65366
```

There are 65366 unique `subject_id`.

```
icustays_tble_count <- icustays_tble |>
  group_by(subject_id) |>
  summarise(num_stays = n()) |>
  ungroup()

ggplot(icustays_tble_count, aes(x = num_stays)) +
  geom_histogram(binwidth = 1) +
  labs(
    title = "Distribution of number of ICU Stays per Subject",
```

```
    x = "Number of ICU Stays",
    y = "Count"
  ) +
  theme_minimal()
```




```
rm(ADT, chartevents, con, d_icd_diagnoses, d_icd_procedures,
   d_items, diagnoses, events, icustays_tble_count, Lab, labevents,
   patient, patient_info, proc, procedures, top_3_diagnoses, transfers,
   vitals, admissions)
gc()
```

```
         used  (Mb) gc trigger  (Mb) max used  (Mb)
Ncells 2045054 109.3    3395975 181.4  3395975 181.4
Vcells 4776982  36.5   23624834 180.3 36913802 281.7
```

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:

```
# Read info from admissions.csv.gz
file_path <- "~/mimic/hosp/admissions.csv.gz"
admissions_tble <- fread(file_path) |> 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:

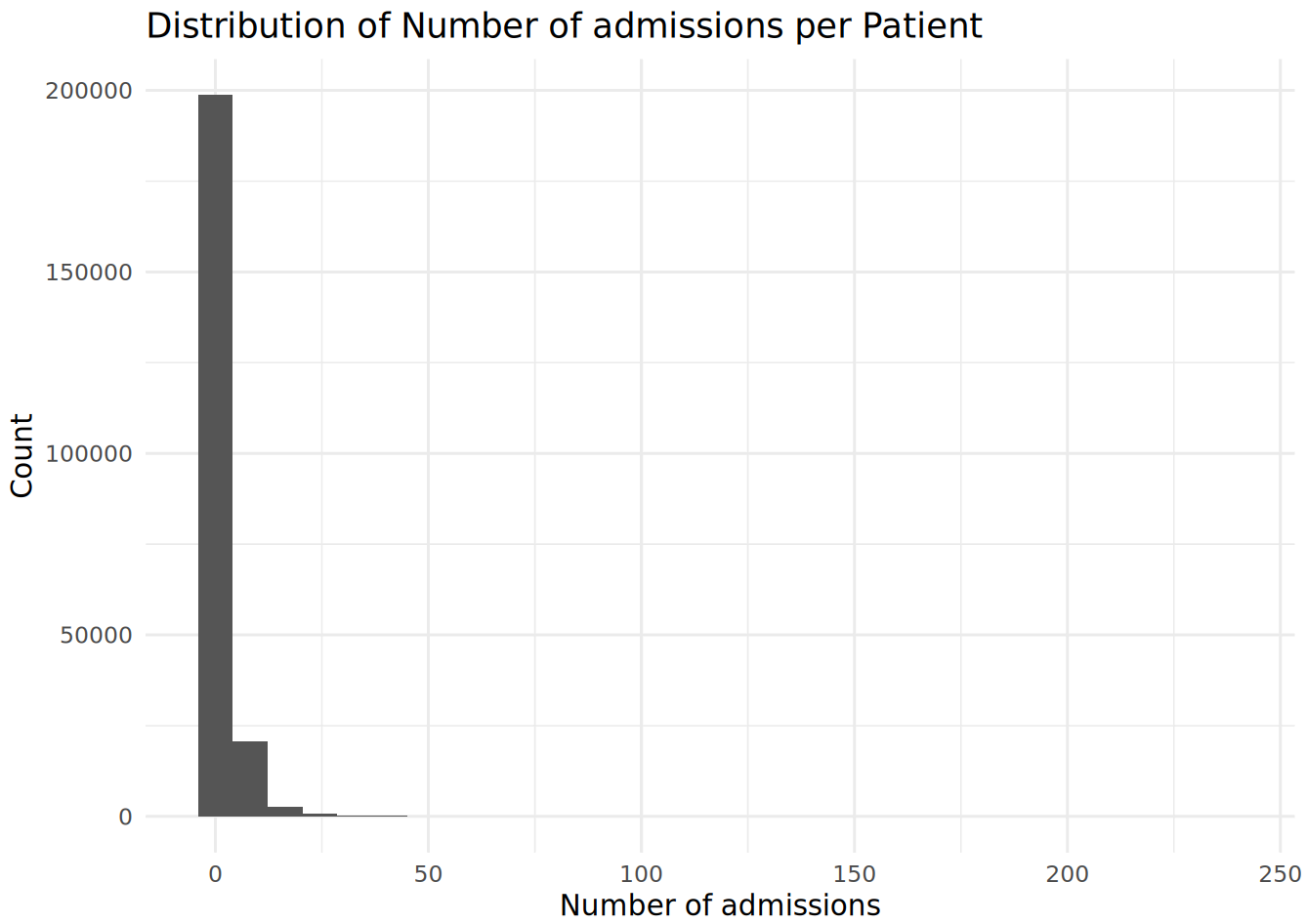
```
admissions_tble_count <- admissions_tble |>
  group_by(subject_id) |>
  summarise(num_admissions = n()) |>
  ungroup()

admissions_tble <- admissions_tble |>
  mutate(admittime = ymd_hms(admittime),
         dischtime = ymd_hms(dischtime),
         admission_hour = hour(admittime),
         admission_minute = minute(admittime),
         los_days = as.numeric(dischtime - admittime, units = "days"),
         los_hours = as.numeric(dischtime - admittime, units = "hours"))
```

```
Warning: There were 2 warnings in `mutate()`.
The first warning was:
ℹ In argument: `admittime = ymd_hms(admittime)`.
Caused by warning:
!  18121 failed to parse.
ℹ Run `dplyr::last_dplyr_warnings()` to see the 1 remaining warning.
```

Visualizations of information:

```
# number of admissions per patient
ggplot(admissions_tble_count, aes(x = num_admissions)) +
  geom_histogram() +
  labs(
    title = "Distribution of Number of admissions per Patient",
    x = "Number of admissions",
    y = "Count"
  ) +
  theme_minimal()
```

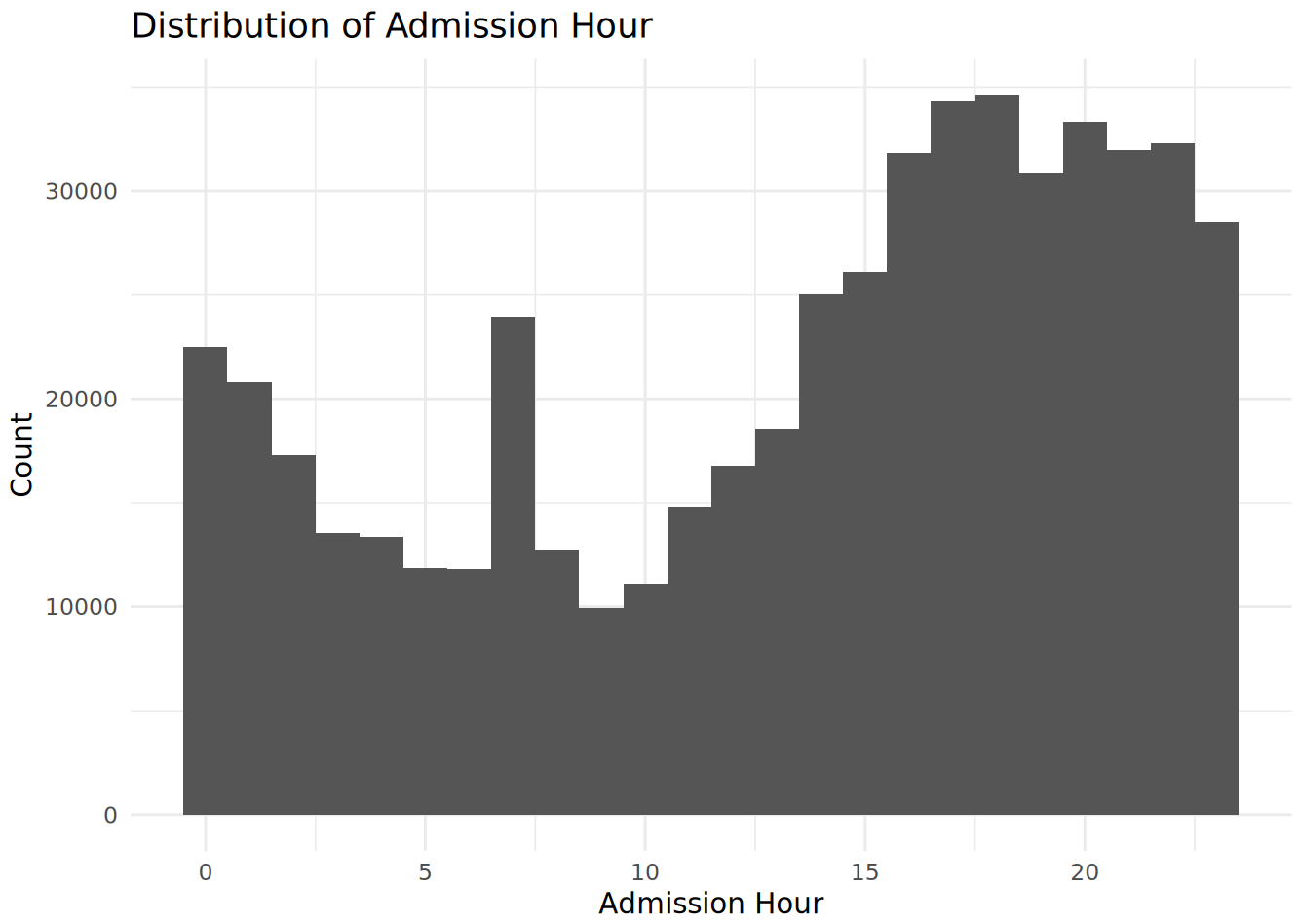
```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The number of admission per patient is right skewed.

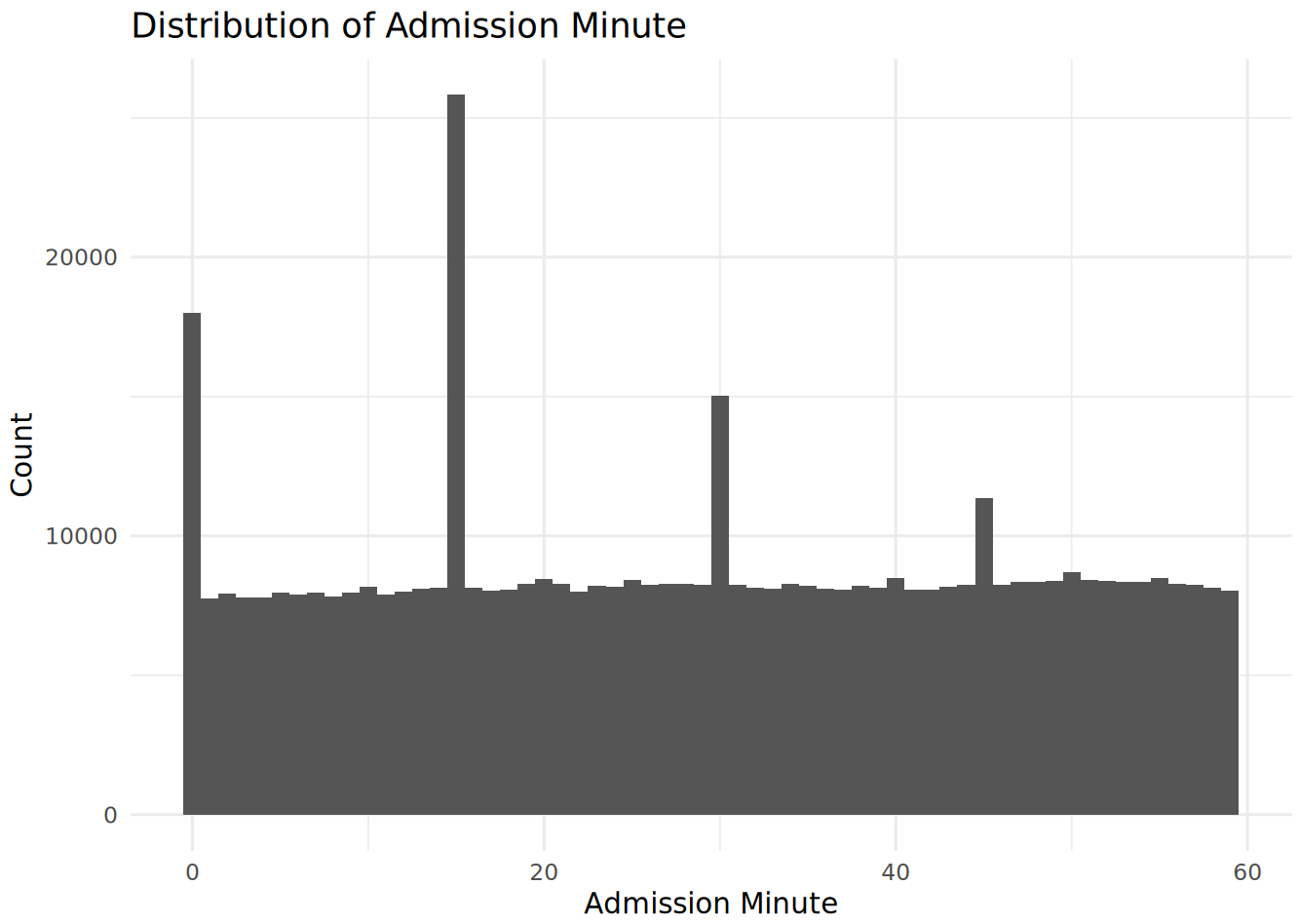
```
# admission hour
ggplot(admissions_tble, aes(x = admission_hour)) +
  geom_histogram(bins = 24) +
  labs(
    title = "Distribution of Admission Hour",
    x = "Admission Hour",
    y = "Count"
  ) +
  theme_minimal()
```

```
Warning: Removed 18121 rows containing non-finite outside the scale range
(`stat_bin()`).
```



In general, admissions decrease from 12 AM to 10 AM and then rise again until midnight, peaking around 6 PM. Interestingly, there is a noticeable spike in admissions at 7 AM compared to the surrounding hours.

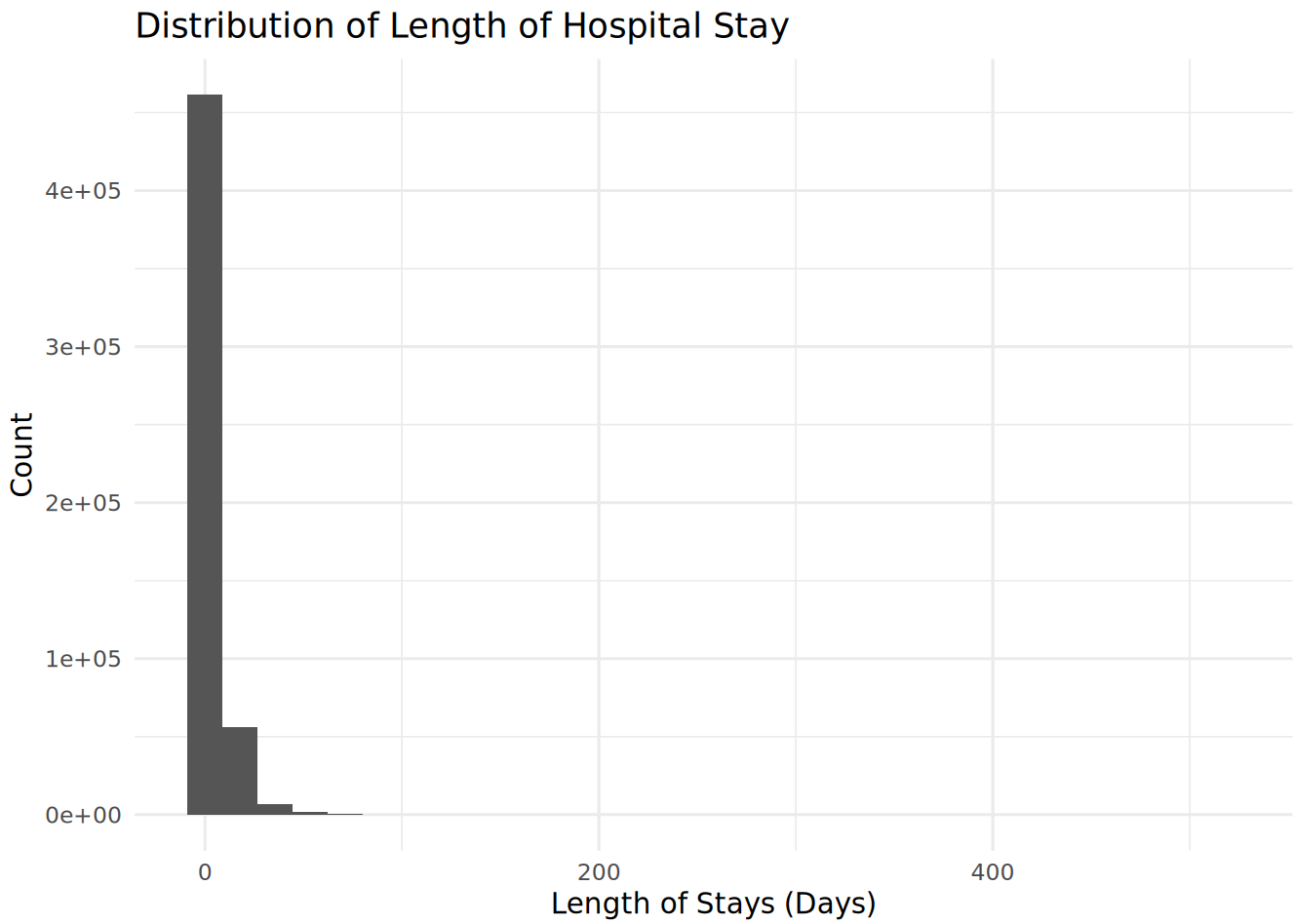
```
# admission minute
ggplot(admissions_tble, aes(x = admission_minute)) +
  geom_histogram(bins = 60) +
  labs(
    title = "Distribution of Admission Minute",
    x = "Admission Minute",
    y = "Count"
  ) +
  theme_minimal()
```



In general, admission minutes are uniformly distributed. However, there are noticeable spikes at whole hours, as well as at the 15-minute, 30-minute, and 45-minute marks.

```
# length of hospital stay
ggplot(admissions_tble, aes(x = los_days)) +
  geom_histogram() +
  labs(
    title = "Distribution of Length of Hospital Stay",
    x = "Length of Stays (Days)",
    y = "Count"
  ) +
  theme_minimal()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Length of stays distribution is right skewed. No unusual pattern identified.

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

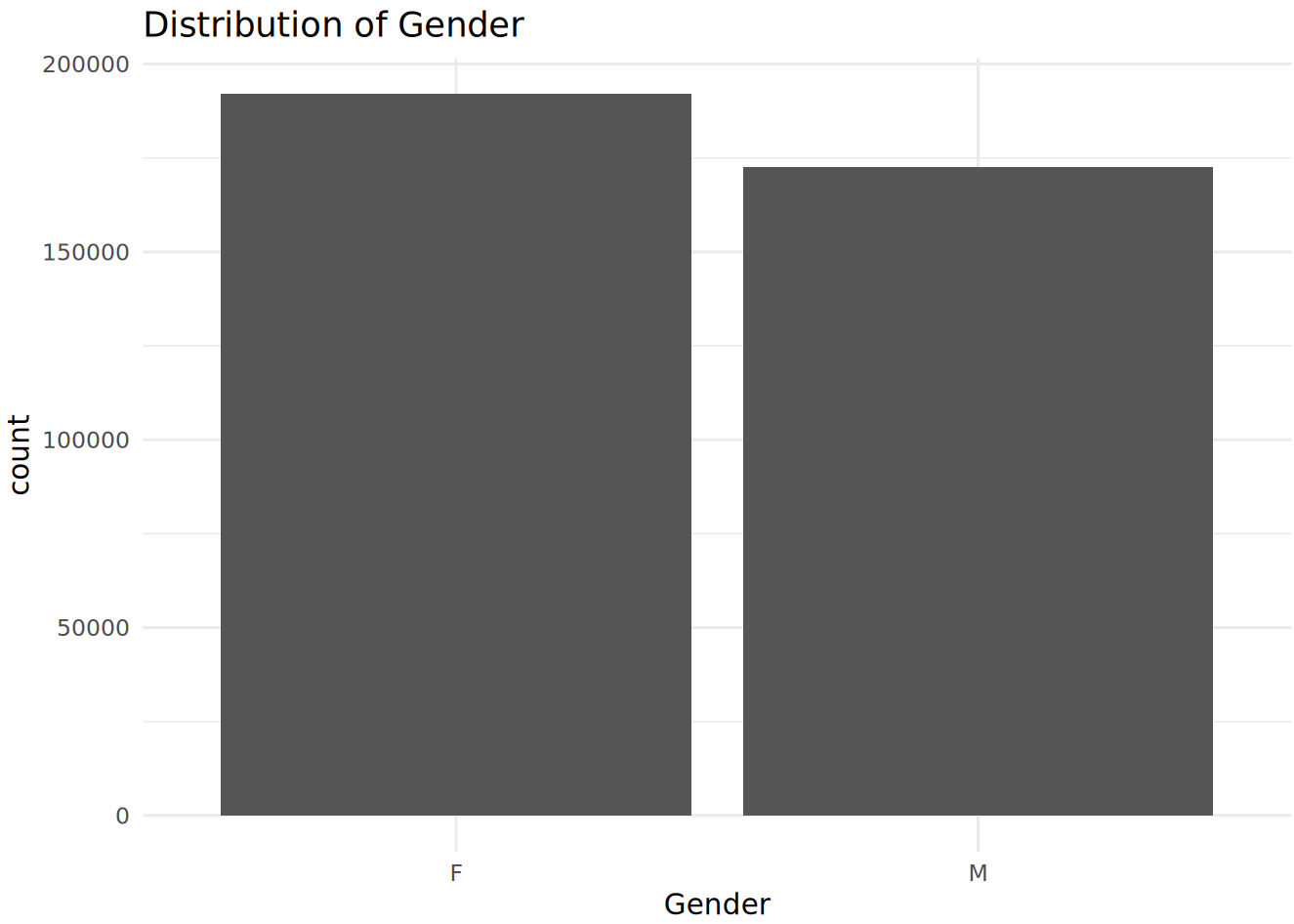
Import `patients.csv.gz` (https://mimic.mit.edu/docs/iv/modules/hosp/patients/) as a tibble `patients_tble`.

```
# import patients.csv.gz
file_path <- "~/mimic/hosp/patients.csv.gz"
patients_tble <- fread(file_path) |> as_tibble()
```

Q4.2 Summary and visualization

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

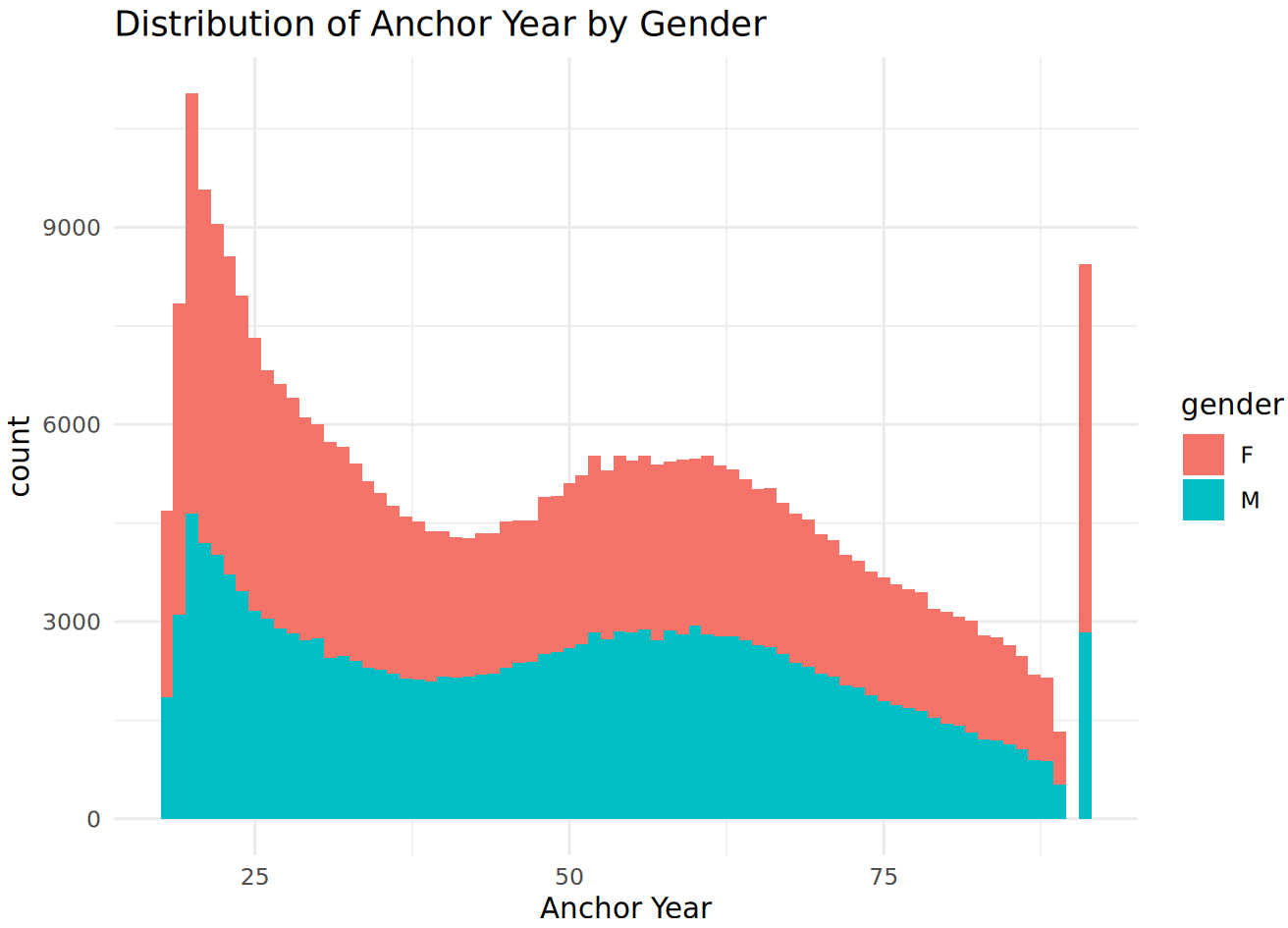
```
# summarize gender
ggplot(patients_tble, aes(x = gender)) +
  geom_bar() +
  labs(
    title = "Distribution of Gender",
    x = "Gender"
  ) +
  theme_minimal()
```



There are slightly more female patients than male patients.

```
# summarize anchor_age
ggplot(patients_tble) +
```

```
  geom_histogram(aes(x = anchor_age, fill = gender), binwidth = 1) +
  labs(
    title = "Distribution of Anchor Year by Gender",
    x = "Anchor Year"
  ) +
  theme_minimal()
```



The distribution of Anchor Age peaks in the early years, then decreases from 25 to 40, followed by a slight increase before steadily declining from 40 to 80 years. Interestingly, there is an outlier with an extremely high value for both male and female patients.

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:

```
itemids <- c(50912, 50971, 50983, 50902, 50882, 51221, 51301, 50931)

file_path <- "~/mimic/hosp/d_labitems.csv.gz"
d_labitems <- fread(file_path) |>
  filter(itemid %in% itemids) |>
```

```
  select(itemid, label)
d_labitems
```

```
   itemid             label
    <int>            <char>
1:  50882       Bicarbonate
2:  50902          Chloride
3:  50912        Creatinine
4:  50931           Glucose
5:  50971         Potassium
6:  50983            Sodium
7:  51221        Hematocrit
8:  51301 White Blood Cells
```

```
rename_vec <- setNames(d_labitems$label, d_labitems$itemid)
```

```
labevents_tble <- open_dataset("labevents_pq", format = "parquet") |>
  to_duckdb() |>
  # only keep information needed from labevents
  select(subject_id, itemid, storetime, valuenum) |>
  # filter itemid
  filter(itemid %in% itemids) |>
  # merge with icustays to get stay information
  left_join(
    select(icustays_tble, subject_id, stay_id, intime),
    by = c("subject_id"),
    copy = TRUE # copy icustay table from memory to database
  ) |>
  # only keep labevents before the ICU stay
  filter(storetime < intime) |>
  # for each patient, icu stay, itemid combination
  group_by(subject_id, stay_id, itemid) |>
  # only keep the most recent labevent
  slice_max(storetime, n = 1) |>
  # discard storetime and intime
  select(-storetime, -intime) |>
  ungroup() |>
  # record lab items to be columns
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  # rename column names from itemid to label
  rename_at(
    vars(names(rename_vec)), ~rename_vec[.]
  ) |>
  rename_with(tolower) |>
  rename(wbc = `white blood cells`) |>
  arrange(subject_id, stay_id) |>
  # reorder column
  select(
    subject_id, stay_id, bicarbonate, chloride, creatinine,
```

```
    glucose, potassium, sodium, hematocrit, wbc) |>
  collect() |>
  as_tibble()
```

```
labevents_tble
```

```
# A tibble: 88,086 × 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 88,076 more rows
# 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.

```
> 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.

Solution:

```
itemids <- c(220045, 220179, 220180, 223761, 220210)
file_path <- "~/mimic/icu/d_items.csv.gz"
d_items <- fread(file_path) |>
  filter(itemid %in% itemids) |>
  select(itemid, label)
d_items <- d_items |>
  mutate(label = tolower(gsub(" ", "_", label)))
d_items
```

```
   itemid                                label
    <int>                               <char>
1: 220045                           heart_rate
2: 220179  non_invasive_blood_pressure_systolic
3: 220180 non_invasive_blood_pressure_diastolic
```

```
4: 220210                     respiratory_rate
5: 223761               temperature_fahrenheit
```

```
rename_vec <- setNames(d_items$label, d_items$itemid)
```

```
chartevents_tble <- open_dataset("chartevents_pq", format = "parquet") |>
  to_duckdb() |>
  # only keep information needed for chartevents
  select(subject_id, itemid, storetime, valuenum) |>
  # filter itemid
  filter(itemid %in% itemids) |>
  # merge with icustays to get stay information
  left_join(
    select(icustays_tble, subject_id, stay_id, intime, outtime),
    by = c("subject_id"),
    copy = TRUE
  ) |>
  # only keep chartevents within the ICU stay
  filter(storetime >= intime & storetime <= outtime) |>
  # group by each patient, icu stay, itemid combination
  group_by(subject_id, stay_id, itemid, storetime) |>
  # get the average for measurement with same storetime
  mutate(valuenum = mean(valuenum, na.rm = TRUE)) |>
  ungroup() |>
  group_by(subject_id, stay_id, itemid) |>
  # only keep the first vital measurement
  slice_min(storetime, n = 1) |>
  # discard storetime, intime, and outtime
  select(-storetime, -intime, -outtime) |>
  ungroup() |>
  # record chart items to be columns
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  collect() |>
  as_tibble()
```

```
chartevents_tble <- chartevents_tble |>
  rename_at(
    vars(names(rename_vec)),  # Select columns to rename
    ~ rename_vec[.]  # Apply renaming function
  ) |>
  arrange(subject_id, stay_id) |>
  select(
    subject_id, stay_id, heart_rate, non_invasive_blood_pressure_diastolic,
    non_invasive_blood_pressure_systolic, respiratory_rate,
    temperature_fahrenheit)
```

```
chartevents_tble
```

```
# A tibble: 94,364 × 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 94,354 more rows
# 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.

```
> 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
```

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)
```

```
mimic_icu_cohort
```

```
# A tibble: 94,458 × 45
   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 94,448 more rows
# ℹ 39 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>, admission_hour <int>, admission_minute <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)
- 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:

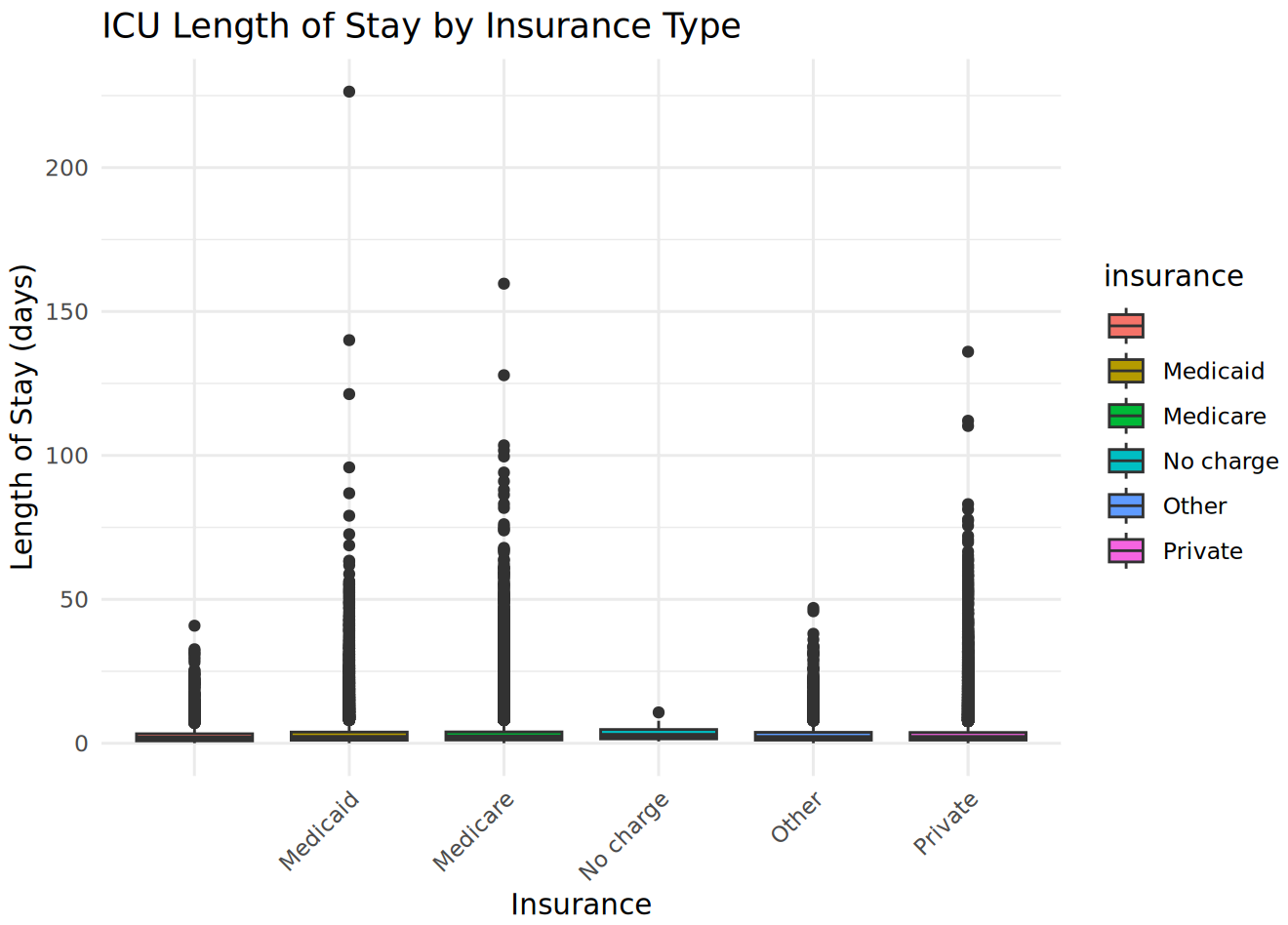
Length of ICU stay vs. demographic variables

```
# Length of ICU stay vs. Race
los_race_summary <- mimic_icu_cohort %>%
  group_by(race) %>%
  summarize(
    count = n(),
```

```
    mean_los = mean(los, na.rm = TRUE),
    median_los = median(los, na.rm = TRUE),
    IQR_los = IQR(los, na.rm = TRUE),
    sd_los = sd(los, na.rm = TRUE),
    .groups = "drop"
  )
los_race_summary
```

```
# A tibble: 33 × 6
   race                          count mean_los median_los IQR_los sd_los
   <chr>                         <int>    <dbl>      <dbl>   <dbl>  <dbl>
 1 AMERICAN INDIAN/ALASKA NATIVE   198     4.31       2.08    3.61   6.48
 2 ASIAN                          1095     3.56       1.92    2.70   4.95
 3 ASIAN - ASIAN INDIAN            248     4.08       1.90    2.73   7.28
 4 ASIAN - CHINESE                1062     3.59       1.89    2.52   5.47
 5 ASIAN - KOREAN                   73     4.44       2.25    2.64   7.42
 6 ASIAN - SOUTH EAST ASIAN        408     3.45       1.86    2.22   5.72
 7 BLACK/AFRICAN                   431     4.01       2.08    3.12   5.80
 8 BLACK/AFRICAN AMERICAN         8677     3.54       1.90    2.69   5.42
 9 BLACK/CAPE VERDEAN              656     3.67       1.83    2.67   6.32
10 BLACK/CARIBBEAN ISLAND          621     4.34       2.04    2.87   6.97
# ℹ 23 more rows
```

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



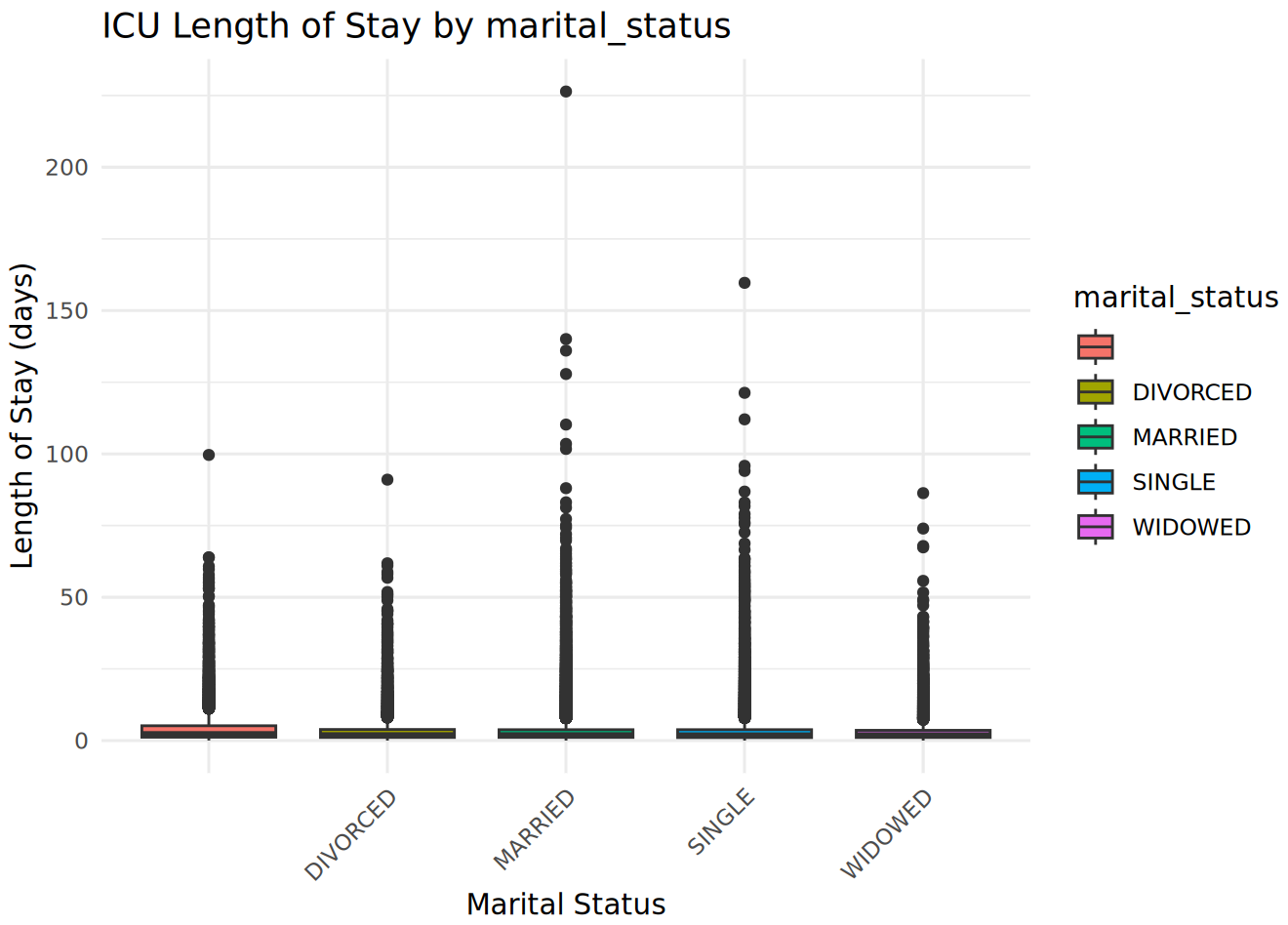

```
los_insurance_summary <- mimic_icu_cohort %>%
  group_by(insurance) %>%
  summarize(
    count = n(),
    mean_los = mean(los, na.rm = TRUE),
    median_los = median(los, na.rm = TRUE),
    IQR_los = IQR(los, na.rm = TRUE),
    sd_los = sd(los, na.rm = TRUE),
    .groups = "drop"
  )
los_insurance_summary
```

```
# A tibble: 6 × 6
  insurance   count mean_los median_los IQR_los sd_los
  <chr>       <int>    <dbl>      <dbl>   <dbl>  <dbl>
1 ""           1523     3.21       1.65    2.47   4.52
2 "Medicaid"  14240     3.79       1.90    2.82   6.21
3 "Medicare"  51819     3.60       2.03    2.81   5.10
4 "No charge"     8     3.87       2.60    3.26   3.55
5 "Other"      2328     3.39       1.86    2.71   4.56
6 "Private"   24540     3.64       1.88    2.64   5.64
```

```
# Length of ICU stay vs. marital_status
los_marital_status_summary <- mimic_icu_cohort %>%
  mutate(marital_status = ifelse(
    marital_status == "", "OTHER", marital_status)) %>%
  group_by(marital_status) %>%
  summarize(
    count = n(),
    mean_los = mean(los, na.rm = TRUE),
    median_los = median(los, na.rm = TRUE),
    IQR_los = IQR(los, na.rm = TRUE),
    sd_los = sd(los, na.rm = TRUE),
    .groups = "drop"
  )
los_marital_status_summary
```

```
# A tibble: 5 × 6
  marital_status count mean_los median_los IQR_los sd_los
  <chr>          <int>    <dbl>      <dbl>   <dbl>  <dbl>
1 DIVORCED        6932     3.58       1.95    2.78   5.16
2 MARRIED        41907     3.59       1.97    2.68   5.44
3 OTHER           7761     4.64       2.33    3.98   6.40
4 SINGLE         26785     3.59       1.91    2.75   5.49
5 WIDOWED        11073     3.18       1.93    2.50   4.22
```

```
ggplot(mimic_icu_cohort, aes(x = marital_status, y = los,
                             fill = marital_status)) +
  geom_boxplot() +
  labs(
    title = "ICU Length of Stay by marital_status",
    x = "Marital Status",
    y = "Length of Stay (days)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

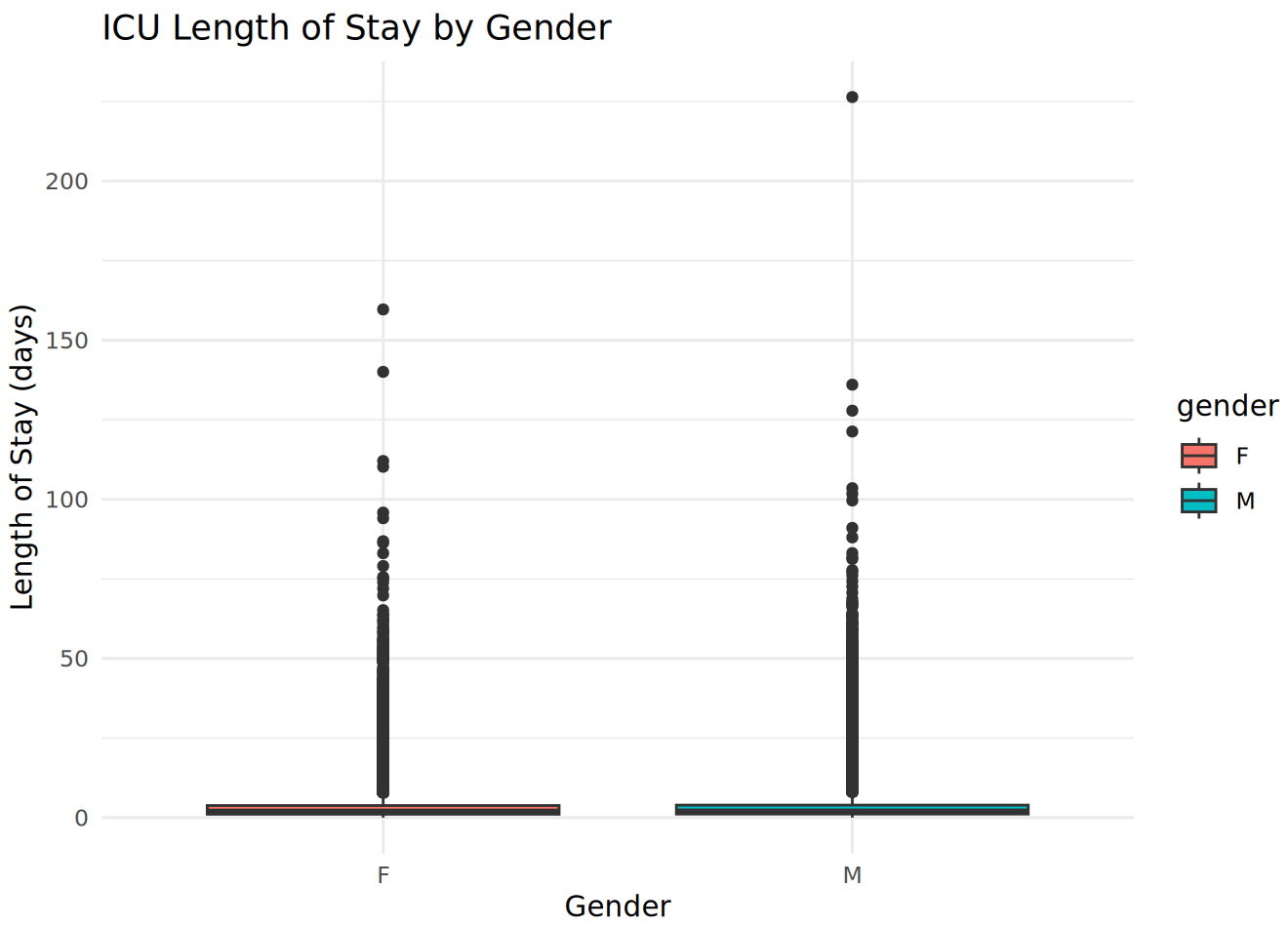



```
# Length of ICU stay vs. gender
los_gender_summary <- mimic_icu_cohort %>%
  group_by(gender) %>%
  summarize(
    count = n(),
    mean_los = mean(los, na.rm = TRUE),
    median_los = median(los, na.rm = TRUE),
    IQR_los = IQR(los, na.rm = TRUE),
    sd_los = sd(los, na.rm = TRUE),
    .groups = "drop"
  )
los_gender_summary
```

```
# A tibble: 2 × 6
  gender count mean_los median_los IQR_los sd_los
  <chr>  <int>    <dbl>      <dbl>   <dbl>  <dbl>
1 F      41583     3.51       1.94    2.73   5.17
2 M      52875     3.72       1.98    2.80   5.58
```

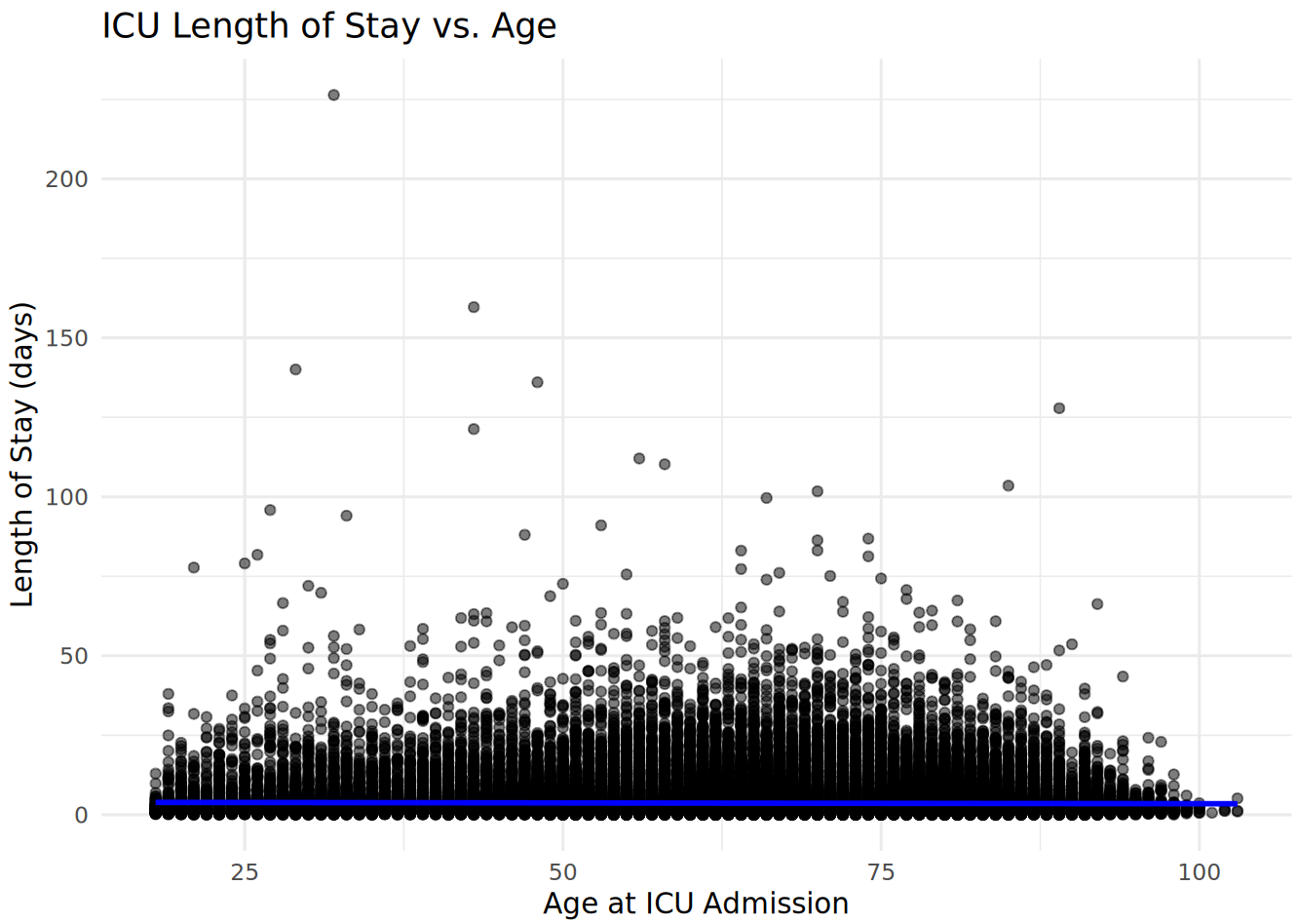
```
# Boxplot for LOS by Gender
ggplot(mimic_icu_cohort, aes(x = gender, y = los, fill = gender)) +
  geom_boxplot() +
  labs(title = "ICU Length of Stay by Gender",
```

```
       x = "Gender", y = "Length of Stay (days)") +
  theme_minimal()
```




```
# Length of ICU stay vs. age at intime
ggplot(mimic_icu_cohort, aes(x = age_intime, y = los)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "ICU Length of Stay vs. Age",
       x = "Age at ICU Admission", y = "Length of Stay (days)") +
  theme_minimal()
```

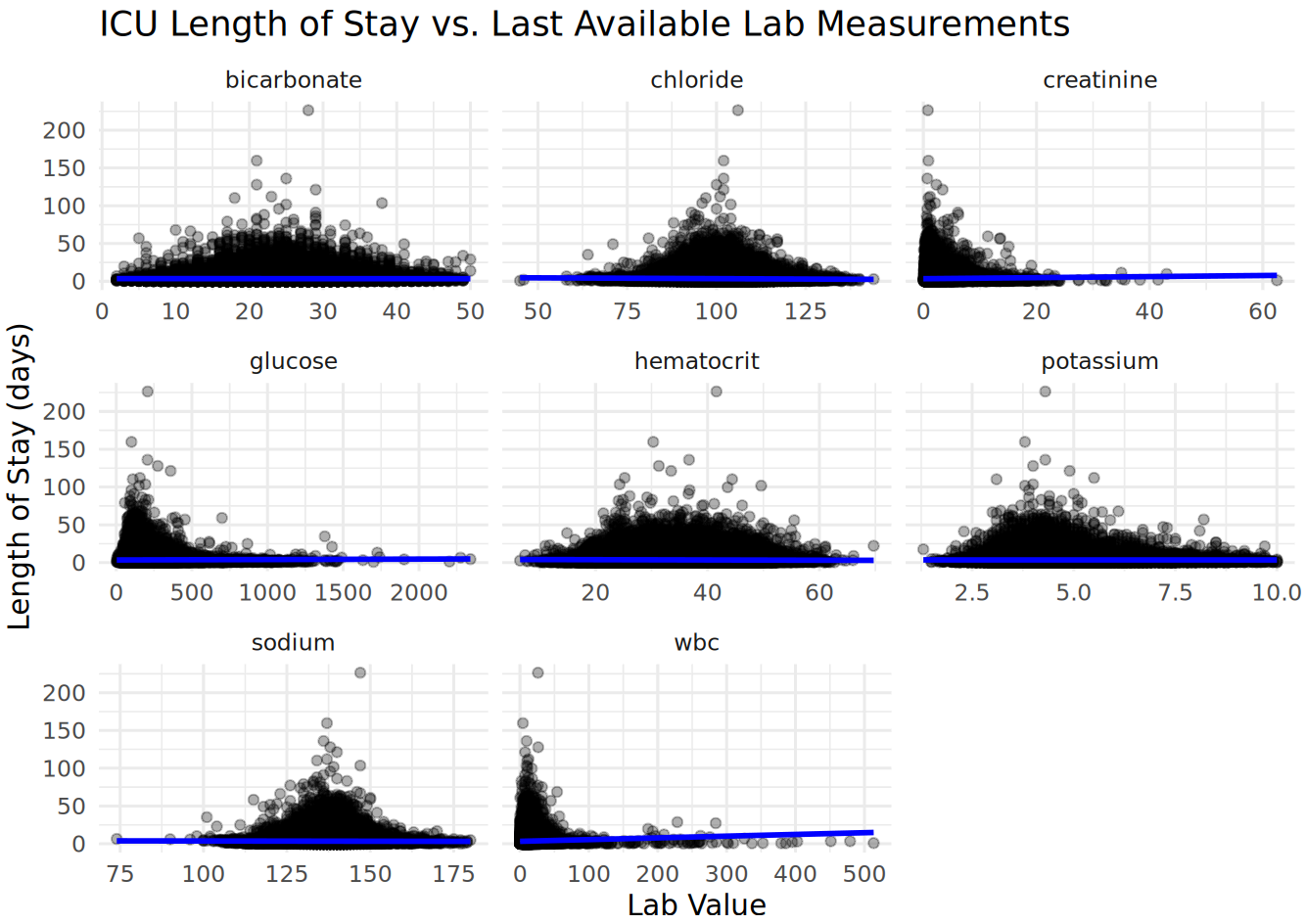
```
`geom_smooth()` using formula = 'y ~ x'
```



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

```
mimic_icu_cohort_long <- mimic_icu_cohort %>%
  pivot_longer(cols = c(bicarbonate, chloride, creatinine, glucose,
                        potassium, sodium, hematocrit, wbc),
               names_to = "lab_name", values_to = "lab_value")

ggplot(mimic_icu_cohort_long, aes(x = lab_value, y = los)) +
  # Scatter plot with transparency to reduce overplotting
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  # Separate plots for each lab item
  facet_wrap(~lab_name, scales = "free_x") +
  labs(title = "ICU Length of Stay vs. Last Available Lab Measurements",
       x = "Lab Value",
       y = "Length of Stay (days)") +
  theme_minimal()
```

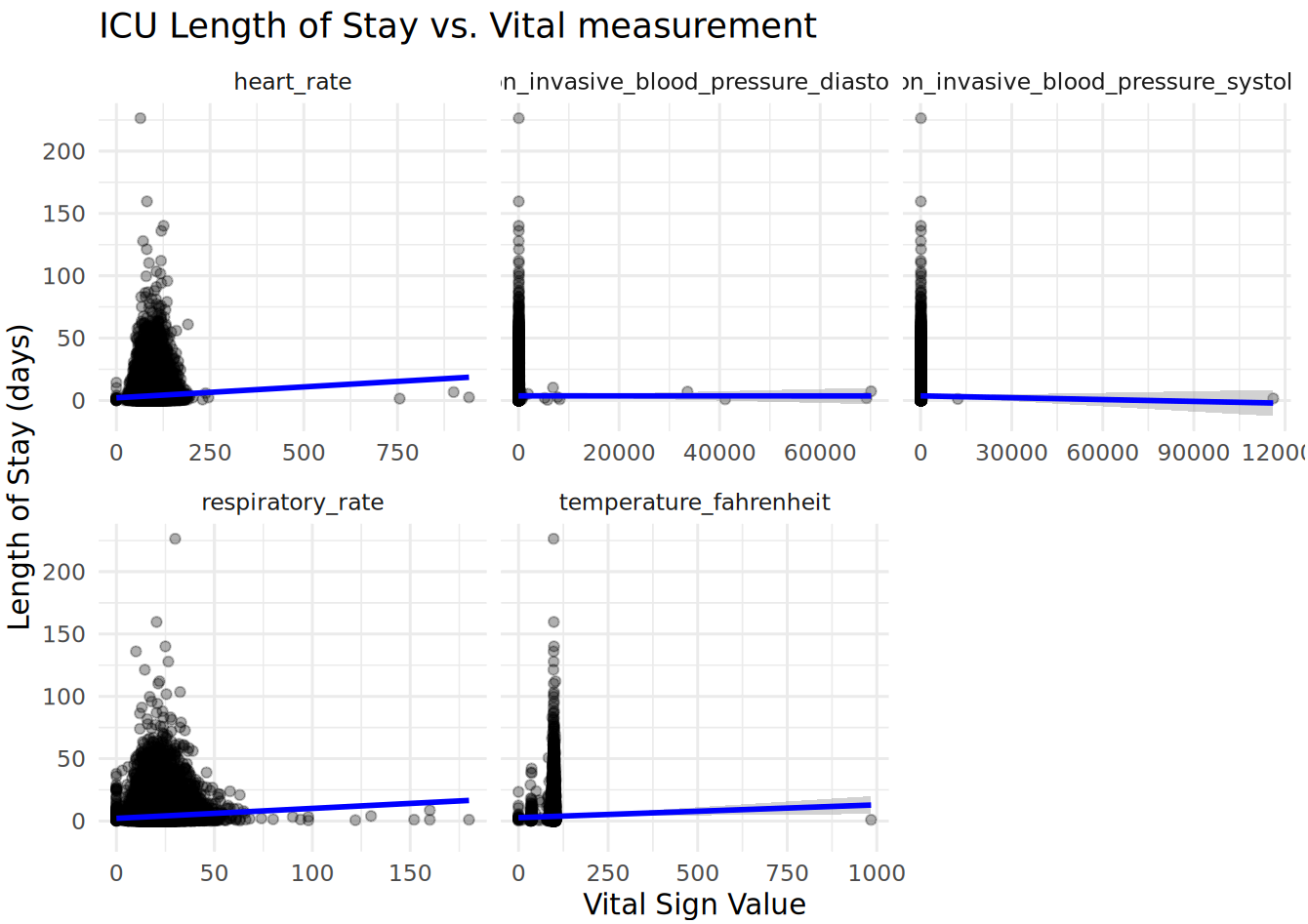
```
`geom_smooth()` using formula = 'y ~ x'
```



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

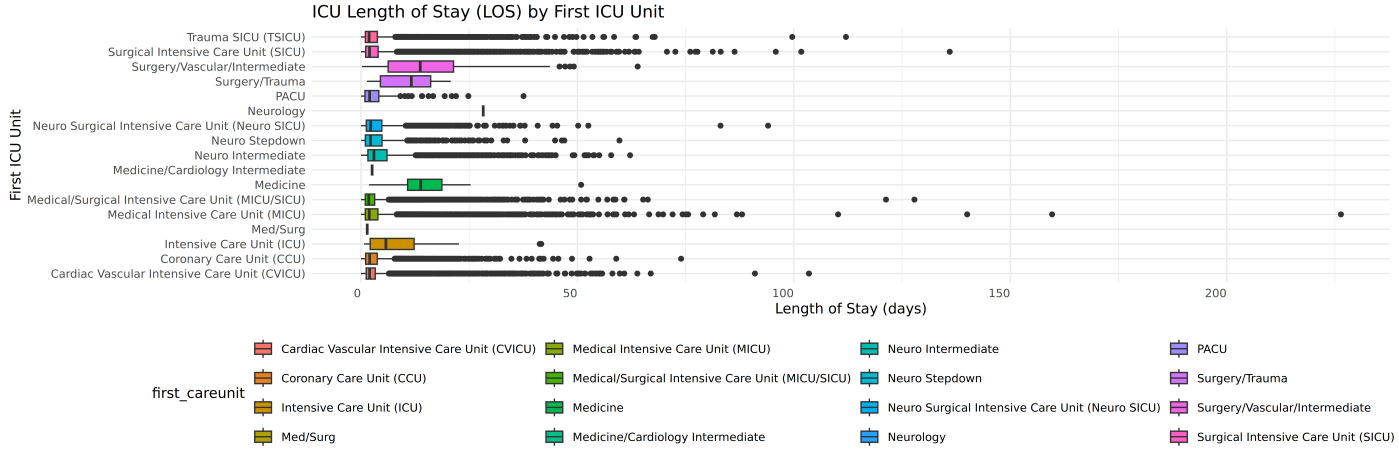
```
mimic_icu_vitals_long <- mimic_icu_cohort %>%
  pivot_longer(cols = c(heart_rate, non_invasive_blood_pressure_systolic,
                        non_invasive_blood_pressure_diastolic,
                        respiratory_rate, temperature_fahrenheit),
               names_to = "vital_sign", values_to = "value")

ggplot(mimic_icu_vitals_long, aes(x = value, y = los)) +
  geom_point(alpha = 0.3) +  # Scatter points with transparency
  geom_smooth(method = "lm", color = "blue", se = TRUE) +
  facet_wrap(~vital_sign, scales = "free_x") +
  labs(title = "ICU Length of Stay vs. Vital measurement",
       x = "Vital Sign Value",
       y = "Length of Stay (days)") +
  theme_minimal()
```

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



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 Length of Stay (LOS) by First ICU Unit",
       x = "First ICU Unit",
       y = "Length of Stay (days)") +
  theme_minimal() +
  theme(
    axis.text.x = element_text(hjust = 1),
    legend.position = "bottom") +
  coord_flip()
```


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