Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

Ningke Zhang 705834790

Display machine information for reproducibility:

```
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS Sequoia 15.3.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  L

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

time zone: America/Los_Angeles
tzcode source: internal

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

loaded via a namespace (and not attached):
 [1] compiler_4.4.2    fastmap_1.2.0     cli_3.6.3         tools_4.4.2
 [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29
 [9] knitr_1.49        jsonlite_1.8.9    xfun_0.50         digest_0.6.37
[13] rlang_1.1.4       evaluate_1.0.1
```

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 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2

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

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

Display your machine memory.

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

```
Totalram:  16.000 GiB
Freeram:    6.262 GiB
```

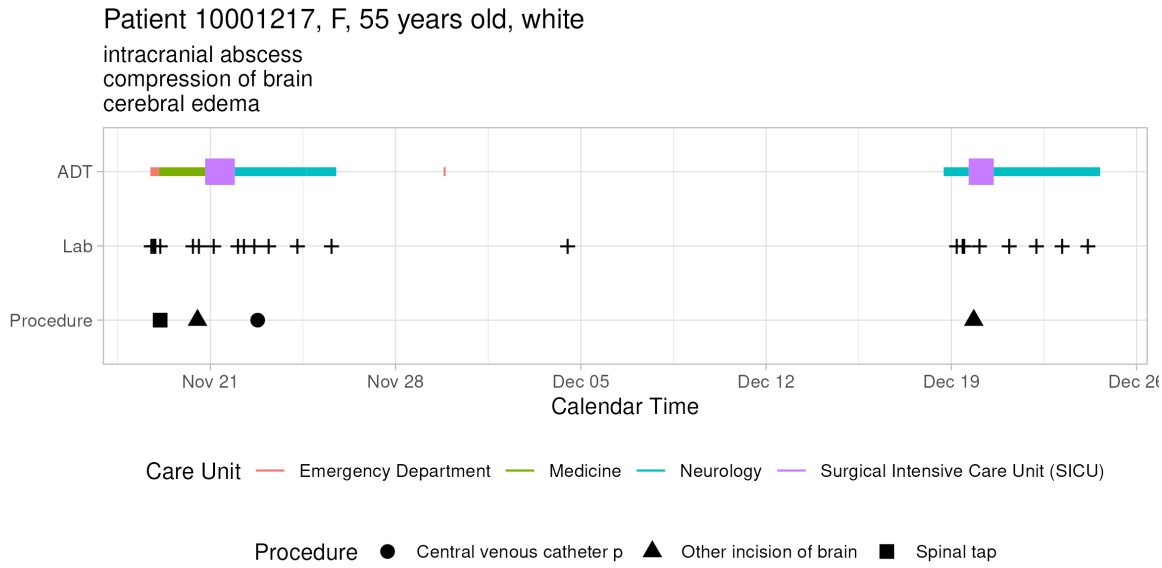
In this exercise, we use tidyverse (ggplot2, dplyr, etc) to explore the MIMIC-IV data introduced in homework 1 and to build a cohort of ICU stays.

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

```
gunzip -c ~/mimic/hosp/labevents.csv.gz > labevents.csv
```

```
sid <- 10063848
# Load data
patients <- read_csv("~/mimic/hosp/patients.csv.gz") |>
  filter(subject_id == sid)
```

```
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("~/mimic/hosp/admissions.csv.gz") |>
  filter(subject_id == sid)
```

```
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("~/mimic/hosp/transfers.csv.gz") |>
  filter(subject_id == sid)
```

```
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_icd <- read_csv("~/mimic/hosp/procedures_icd.csv.gz") |>
  filter(subject_id == sid) |>
  left_join(read_csv("~/mimic/hosp/d_icd_procedures.csv.gz",
            show_col_types = FALSE), by = c("icd_code", "icd_version"))
```

```
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_icd <- read_csv("~/mimic/hosp/diagnoses_icd.csv.gz") |>
  filter(subject_id == sid) |>
  left_join(read_csv("~/mimic/hosp/d_icd_diagnoses.csv.gz",
            show_col_types = FALSE), by = c("icd_code", "icd_version"))
```

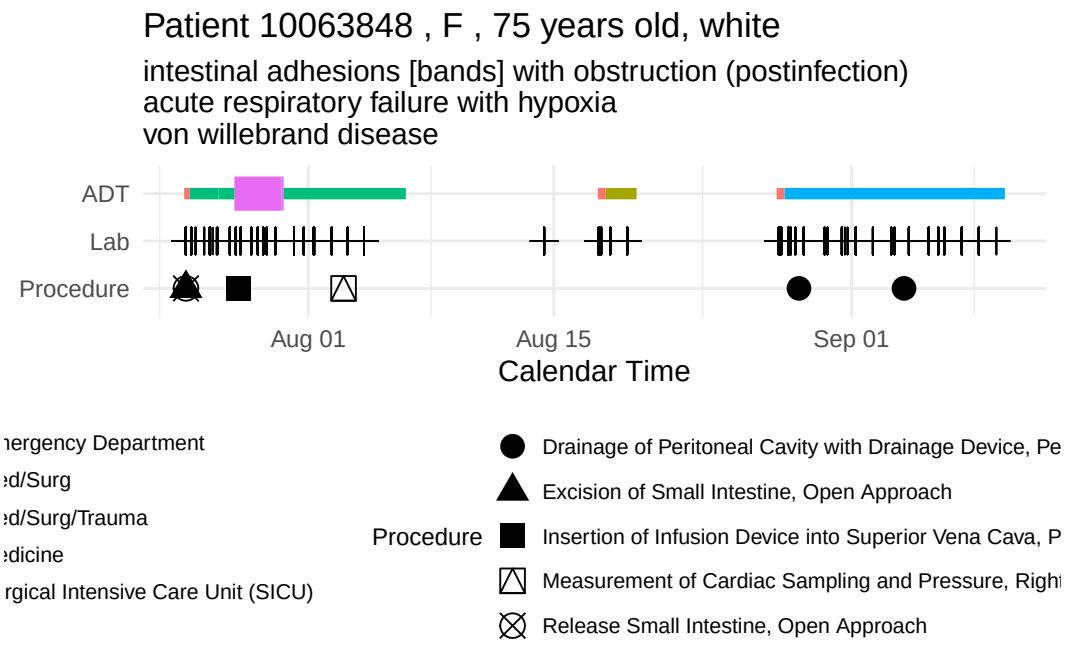
```
Rows: 6364488 Columns: 5
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): icd_code
dbl (4): subject_id, hadm_id, seq_num, icd_version

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.
```

```
labevents_parquet <- "labevents_parquet"
write_dataset(open_dataset("labevents.csv", format = "csv"),
              path = labevents_parquet, format = "parquet")
parquet_ds <- open_dataset(labevents_parquet, format = "parquet")
labevents <- parquet_ds |>
  filter(subject_id == sid) |>
  collect()

# Create plot
transfers <- transfers |>
  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")) |>
  filter(outtime > intime)

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

procedures_icd <- procedures_icd |>
  mutate(chartdate = as.POSIXct(chartdate, format = "%Y-%m-%d", tz = "UTC"))

race <- if("race" %in% colnames(admissions))
  tolower(admissions$race) else "unknown"
unique_careunits <- unique(transfers$careunit)

icu_units <- c("Surgical Intensive Care Unit (SICU)",
               "Medical Intensive Care Unit (MICU)",
```

```
               "Coronary Care Unit (CCU)",
               "Cardiac Vascular Intensive Care Unit (CVICU)",
               "Neuro Surgical Intensive Care Unit (Neuro SICU)")

transfers <- transfers |>
  mutate(is_icu = ifelse(careunit %in% icu_units, "ICU", "Non-ICU"))


ggplot() +
  geom_point(data = procedures_icd, aes(x = chartdate, y = "Procedure",
                                        shape = long_title), size = 4) +
  geom_point(data = labevents, aes(x = charttime, y = "Lab"),
             shape = 3, size = 3) +
  geom_segment(data = transfers,
               aes(x = intime, xend = outtime, y = "ADT",
                   yend = "ADT", color = careunit, size = is_icu)) +
  labs(
    title = paste("Patient", sid, ",", patients$gender, ",",
                  patients$anchor_age, "years old,", race),
    subtitle = str_c(str_to_lower(diagnoses_icd$long_title[1:3]),
                     collapse = "\n"),
    x = "Calendar Time",
    y = NULL
  ) +
  scale_color_manual(name = "Care Unit",
                     values = scales::hue_pal()(length(unique_careunits))) +
  scale_shape_manual(name = "Procedure", values = c(16, 17, 15, 14, 13)) +
  scale_size_manual(values = c("ICU" = 6, "Non-ICU" = 2)) +
  scale_y_discrete(limits = c("Procedure", "Lab", "ADT")) +
  theme_minimal() +
  theme(
    legend.position = "bottom",
    legend.box = "horizontal",
    legend.text = element_text(size = 8),
    legend.title = element_text(size = 9),
    legend.key.size = unit(0.3, "cm")
  ) +
  guides(color = guide_legend(nrow = 6),
         shape = guide_legend(nrow = 5),
         size = "none")
```

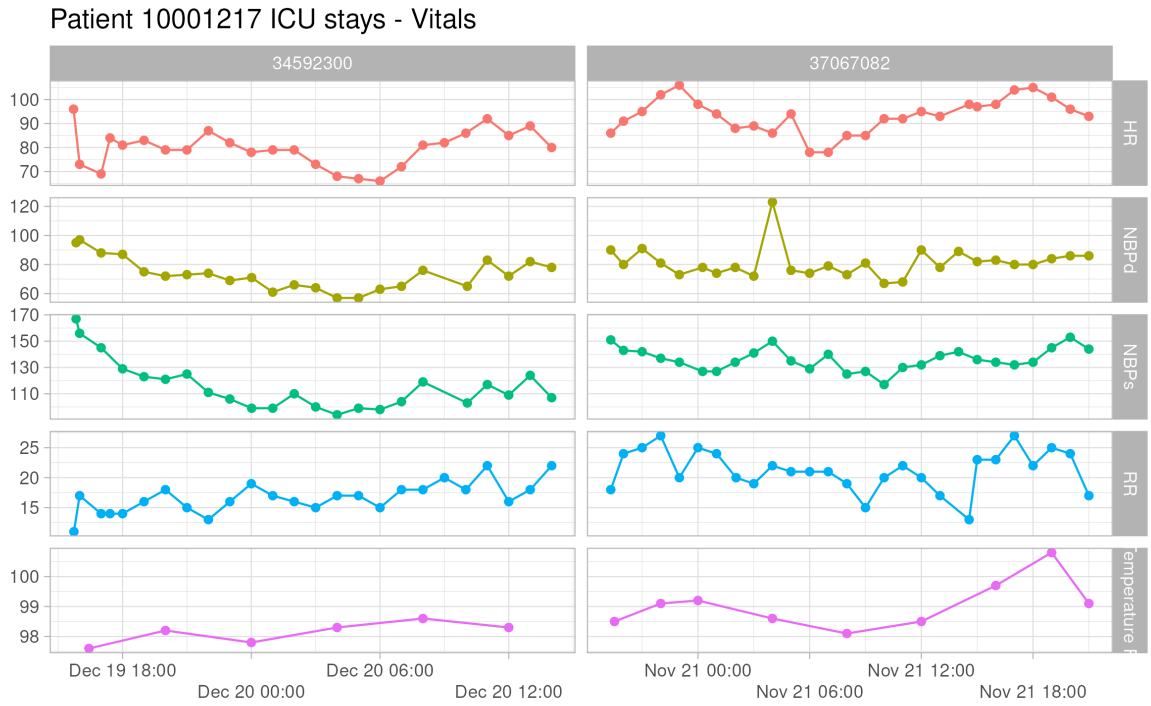
```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
```

```
i Please use `linewidth` instead.
```



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

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

```
# Load data
chartevents_parquet <- "chartevents_parquet"

write_dataset(open_dataset("chartevents.csv", format = "csv"),
              path = chartevents_parquet, format = "parquet")
parquet_ds <- open_dataset(chartevents_parquet, format = "parquet")
chartevents <- parquet_ds |>
  filter(subject_id == sid) |>
  select(subject_id, stay_id, charttime, itemid, value, valuenum) |>
  collect()

head(chartevents, 10)
```

```
# A tibble: 10 x 6
   subject_id  stay_id charttime           itemid value valuenum
```

```
       <int>    <int> <dttm>              <int> <chr>    <dbl>
 1   10063848 31332266 2177-07-28 07:03:00 226531 188.1    188.
 2   10063848 31332266 2177-07-29 14:00:00 220045 93         93
 3   10063848 31332266 2177-07-29 14:00:00 220210 23         23
 4   10063848 31332266 2177-07-29 14:00:00 220277 98         98
 5   10063848 31332266 2177-07-29 14:02:00 220179 97         97
 6   10063848 31332266 2177-07-29 14:02:00 220180 51         51
 7   10063848 31332266 2177-07-29 14:02:00 220181 65         65
 8   10063848 31332266 2177-07-29 03:00:00 225664 105       105
 9   10063848 31332266 2177-07-30 03:00:00 225664 115       115
10   10063848 31332266 2177-07-27 12:00:00 220045 153       153
```

```
d_items <- read_csv("~/mimic/icu/d_items.csv.gz") |>
  select(itemid, label, abbreviation) |>
  print(width = Inf)
```

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

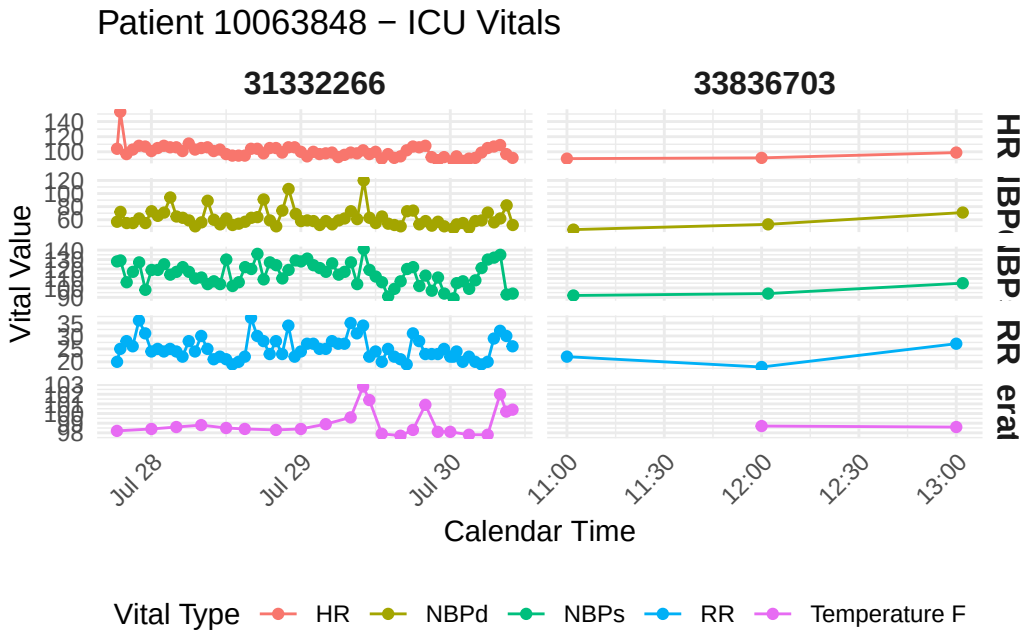
```
# A tibble: 4,095 x 3
   itemid label                                 abbreviation
    <dbl> <chr>                                 <chr>
 1 220001 Problem List                          Problem List
 2 220003 ICU Admission date                    ICU Admission date
 3 220045 Heart Rate                            HR
 4 220046 Heart rate Alarm - High               HR Alarm - High
 5 220047 Heart Rate Alarm - Low                HR Alarm - Low
 6 220048 Heart Rhythm                          Heart Rhythm
 7 220050 Arterial Blood Pressure systolic      ABPs
 8 220051 Arterial Blood Pressure diastolic     ABPd
 9 220052 Arterial Blood Pressure mean          ABPm
10 220056 Arterial Blood Pressure Alarm - Low   ABP Alarm - Low
# i 4,085 more rows
```

```r
chartevents <- chartevents |>
  left_join(d_items, by = "itemid") |>
  filter(abbreviation %in% c("HR", "NBPd", "NBPs", "RR", "Temperature F")) |>
  print(width = Inf)
```

```
# A tibble: 298 x 8
   subject_id  stay_id charttime           itemid value valuenum
        <int>    <int> <dttm>               <dbl> <chr>    <dbl>
 1   10063848 31332266 2177-07-29 14:00:00 220045 93          93
 2   10063848 31332266 2177-07-29 14:00:00 220210 23          23
 3   10063848 31332266 2177-07-29 14:02:00 220179 97          97
 4   10063848 31332266 2177-07-29 14:02:00 220180 51          51
 5   10063848 31332266 2177-07-27 12:00:00 220045 153        153
 6   10063848 31332266 2177-07-27 12:00:00 220210 25          25
 7   10063848 31332266 2177-07-27 12:02:00 220179 129        129
 8   10063848 31332266 2177-07-27 12:02:00 220180 72          72
 9   10063848 31332266 2177-07-27 13:00:00 220045 97          97
10   10063848 31332266 2177-07-27 13:00:00 220210 28          28
   label                                abbreviation
   <chr>                                <chr>
 1 Heart Rate                           HR
 2 Respiratory Rate                     RR
 3 Non Invasive Blood Pressure systolic  NBPs
 4 Non Invasive Blood Pressure diastolic NBPd
 5 Heart Rate                           HR
 6 Respiratory Rate                     RR
 7 Non Invasive Blood Pressure systolic  NBPs
 8 Non Invasive Blood Pressure diastolic NBPd
 9 Heart Rate                           HR
10 Respiratory Rate                     RR
# i 288 more rows
```

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

# Create plot
ggplot(chartevents, aes(x = charttime, y = valuenum, color = abbreviation)) +
  geom_line(size = 0.5) +
  geom_point(size = 1.5, alpha = 1) +
  facet_grid(rows = vars(abbreviation), cols = vars(stay_id),
```

```
            scales = "free") +

  labs(
    title = paste("Patient", sid, "- ICU Vitals"),
    x = "Calendar Time",
    y = "Vital Value",
    color = "Vital Type"
  ) +
  theme_minimal() +
  theme(
    strip.text = element_text(size = 12, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "bottom"
  )
```



Q2. ICU stays

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

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

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

Q2.1 Ingestion

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

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

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(icustays_tble)
```

```
# A tibble: 6 x 8
  subject_id  hadm_id  stay_id first_careunit  last_careunit intime
       <dbl>    <dbl>    <dbl> <chr>           <chr>         <dttm>
1   10000032 29079034 39553978 Medical Intens~ Medical Inte~ 2180-07-23 14:00:00
2   10000690 25860671 37081114 Medical Intens~ Medical Inte~ 2150-11-02 19:37:00
3   10000980 26913865 39765666 Medical Intens~ Medical Inte~ 2189-06-27 08:42:00
4   10001217 24597018 37067082 Surgical Inten~ Surgical Int~ 2157-11-20 19:18:02
5   10001217 27703517 34592300 Surgical Inten~ Surgical Int~ 2157-12-19 15:42:24
6   10001725 25563031 31205490 Medical/Surgic~ Medical/Surg~ 2110-04-11 15:52:22
# i 2 more variables: outtime <dttm>, los <dbl>
```

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** There are 65366 unique `subject_id` in the `icustays` table. Among them, 16242 patients have multiple ICU stays.

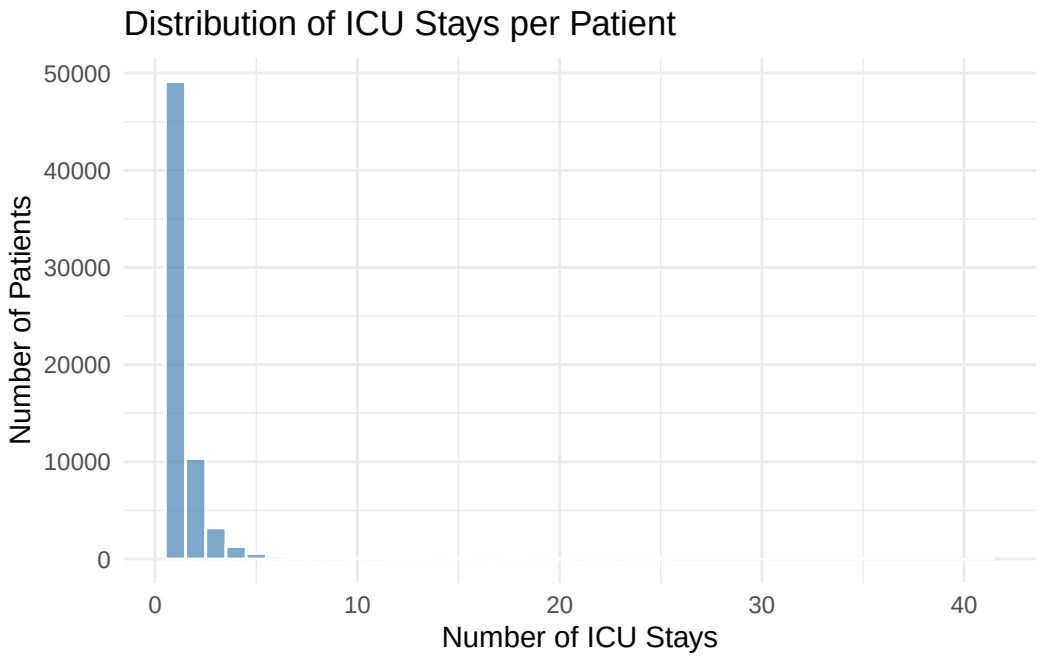
```
# Count unique subject_id
num_unique_subjects <- icustays_tble |>
  distinct(subject_id) |>
  nrow()
print(paste("Number of unique subject_id:", num_unique_subjects))
```

```
[1] "Number of unique subject_id: 65366"
```

```
# Count ICU stays per subject_id
icu_stay_counts <- icustays_tble |>
  group_by(subject_id) |>
  summarise(num_stays = n(), .groups = "drop")
# Check for multiple ICU stays
multi_stay_patients <- icu_stay_counts |>
  filter(num_stays > 1) |>
  nrow()
print(paste("Number of patients with multiple ICU stays:",
            multi_stay_patients))
```

```
[1] "Number of patients with multiple ICU stays: 16242"
```

```
# Create plot
ggplot(icu_stay_counts, aes(x = num_stays)) +
  geom_histogram(binwidth = 1, fill = "steelblue",
                 color = "white", alpha = 0.7) +
  labs(
    title = "Distribution of ICU Stays per Patient",
    x = "Number of ICU Stays",
    y = "Number of Patients"
  ) +
  theme_minimal()
```



Q3. `admissions` data

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

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

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

Q3.1 Ingestion

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

Solution

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

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(admissions_tble)
```

```
# A tibble: 6 x 16
  subject_id hadm_id admittime           dischtime           deathtime
       <dbl>   <dbl> <dttm>              <dttm>              <dttm>
1   10000032  2.26e7 2180-05-06 22:23:00 2180-05-07 17:15:00 NA
2   10000032  2.28e7 2180-06-26 18:27:00 2180-06-27 18:49:00 NA
3   10000032  2.57e7 2180-08-05 23:44:00 2180-08-07 17:50:00 NA
4   10000032  2.91e7 2180-07-23 12:35:00 2180-07-25 17:55:00 NA
5   10000068  2.50e7 2160-03-03 23:16:00 2160-03-04 06:26:00 NA
6   10000084  2.31e7 2160-11-21 01:56:00 2160-11-25 14:52:00 NA
# i 11 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>
```

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,

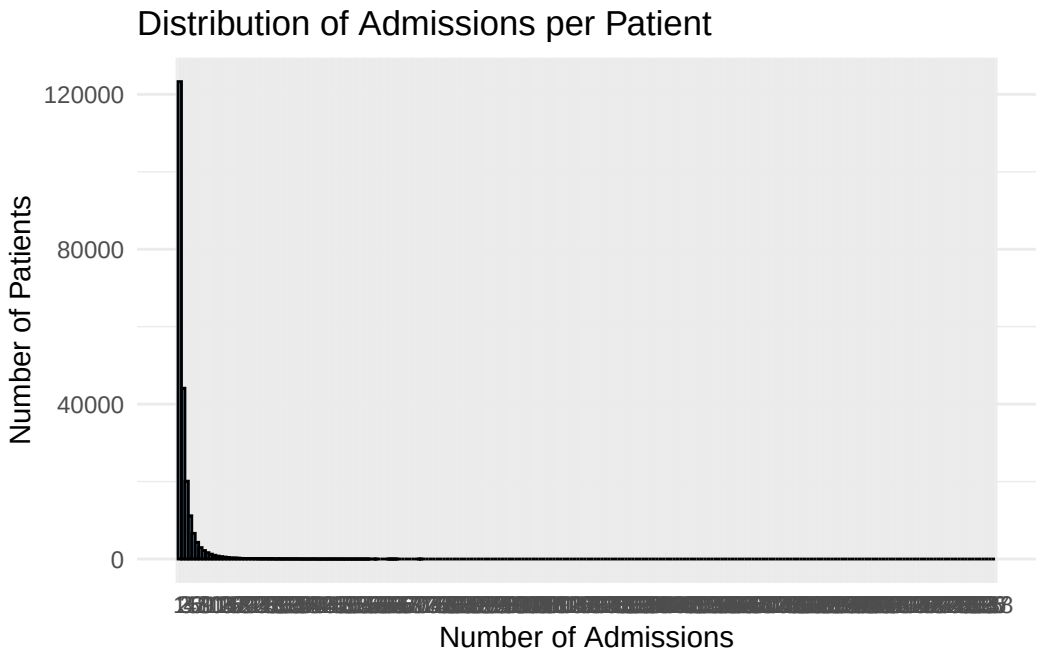
> All dates in the database have been shifted to protect patient confidentiality. Dates will be internally consistent for the same patient, but randomly distributed in the future. Dates of birth which occur in the present time are not true dates of birth. Furthermore, dates of birth which occur before 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 The majority of patients have only one hospital admission, and some patients have multiple readmissions. Peak admissions occur between 3 PM - 6 PM. Notable spikes at midnight and 7 AM may be due to shift changes or scheduled surgeries. There are clear spikes at 00, 15, 30, and 45 minutes. This may be due to rounding or scheduling. Most patients stay 1-3 days in the hospital. There are a few outliers with very long stays, which may be due to critical conditions.

```
# Number of admissions per patient
admission_counts <- admissions_tble |>
  group_by(subject_id) |>
  summarise(num_admissions = n(), .groups = "drop")
summary(admission_counts$num_admissions)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   2.444   3.000 238.000
```

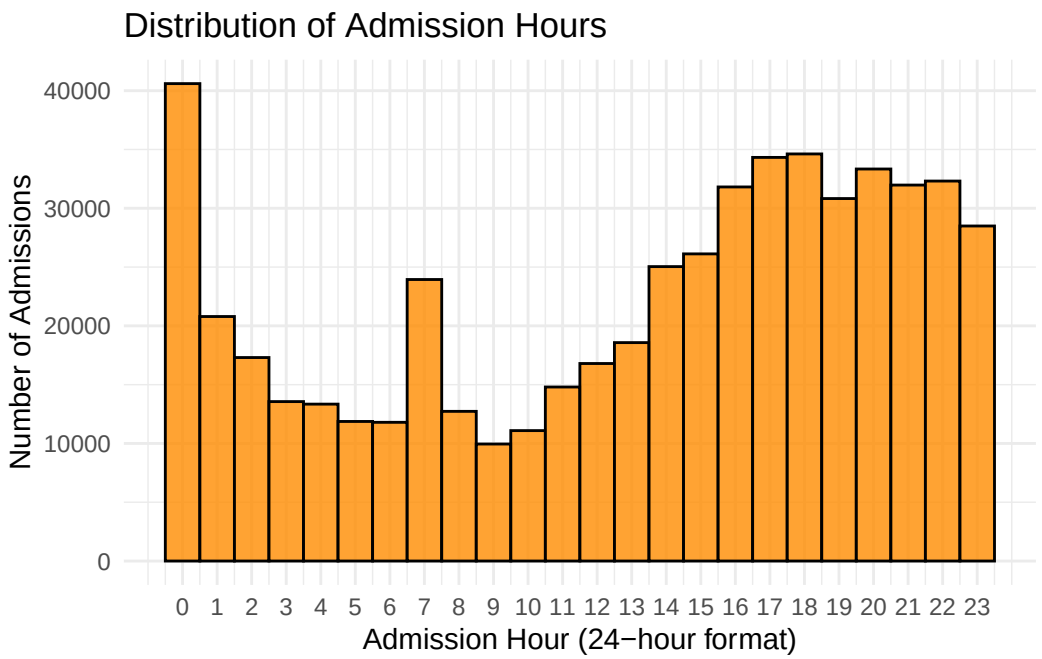
```
ggplot(admission_counts, aes(x = num_admissions)) +
  geom_histogram(binwidth = 1,
                 fill = "steelblue", color = "black", alpha = 0.7) +
  scale_x_continuous(breaks = seq(1, max(admission_counts$num_admissions),
                                  by = 1)) +
  labs(
    title = "Distribution of Admissions per Patient",
    x = "Number of Admissions",
    y = "Number of Patients"
  ) +
  theme_minimal()
```




```
# Admission hour
admissions_tble <- admissions_tble |>
  mutate(admit_hour = lubridate::hour(admittime))
summary(admissions_tble$admit_hour)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    7.00   15.00   13.01   19.00   23.00
```

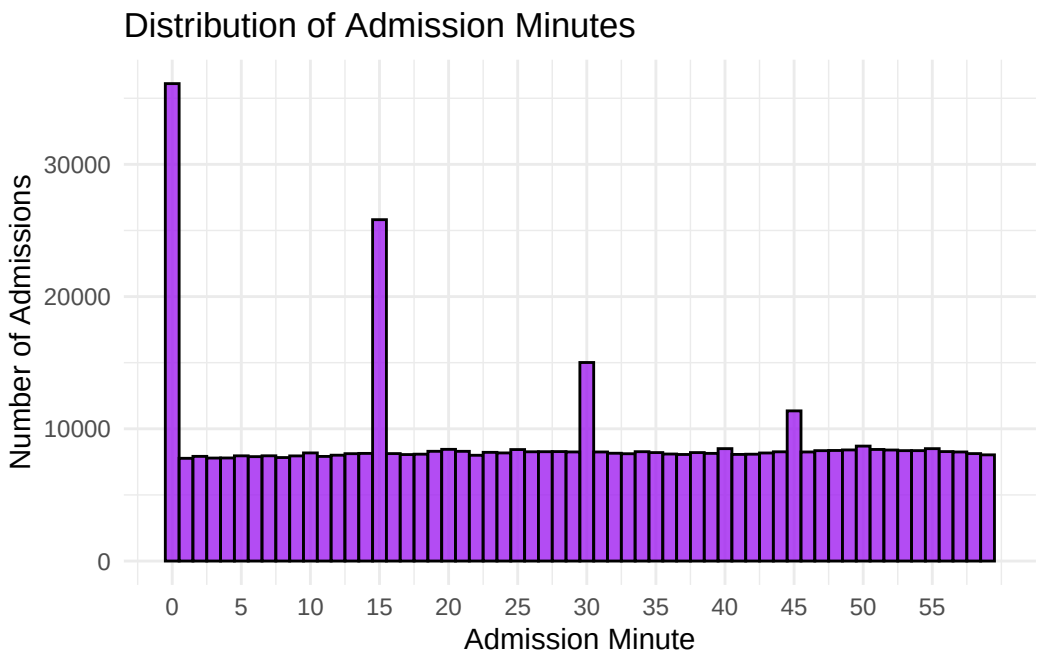
```
ggplot(admissions_tble, aes(x = admit_hour)) +
  geom_histogram(binwidth = 1, fill = "darkorange",
                 color = "black", alpha = 0.8) +
  scale_x_continuous(breaks = seq(0, 23, by = 1)) +
  labs(
    title = "Distribution of Admission Hours",
    x = "Admission Hour (24-hour format)",
    y = "Number of Admissions"
  ) +
  theme_minimal()
```




```
# Admission minute
admissions_tble <- admissions_tble |>
  mutate(admit_minute = lubridate::minute(admittime))
summary(admissions_tble$admit_minute)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   13.00   28.00   27.85   43.00   59.00
```

```
ggplot(admissions_tble, aes(x = admit_minute)) +
  geom_histogram(binwidth = 1, fill = "purple", color = "black", alpha = 0.8) +
  scale_x_continuous(breaks = seq(0, 59, by = 5)) +
  labs(
    title = "Distribution of Admission Minutes",
    x = "Admission Minute",
    y = "Number of Admissions"
  ) +
  theme_minimal()
```



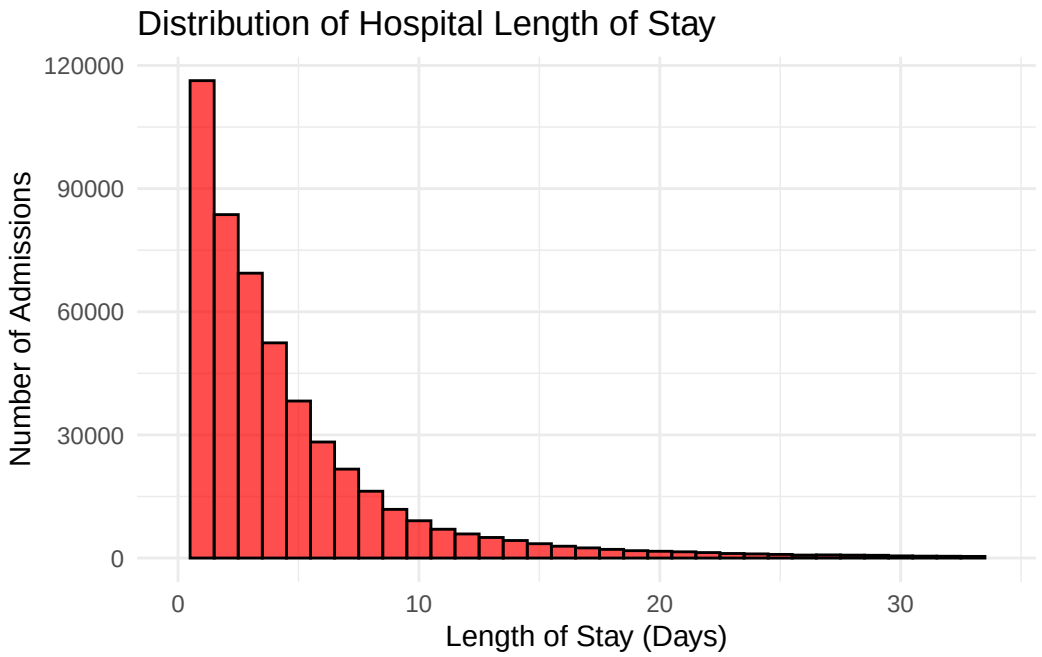

```
# Length of hospital stay
admissions_tble <- admissions_tble |>
  mutate(
    length_of_stay = as.numeric(difftime(dischtime, admittime, units = "days"))
  )
summary(admissions_tble$length_of_stay)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.9451   1.1285   2.8181   4.7617   5.6215 515.5625
```

```
ggplot(admissions_tble, aes(x = length_of_stay)) +
  geom_histogram(binwidth = 1, fill = "red", color = "black", alpha = 0.7) +
  scale_x_continuous(
    limits = c(0, quantile(admissions_tble$length_of_stay, 0.99))) +
  labs(
    title = "Distribution of Hospital Length of Stay",
    x = "Length of Stay (Days)",
    y = "Number of Admissions"
  ) +
  theme_minimal()
```

```
Warning: Removed 5636 rows containing non-finite outside the scale range
```

```
(`stat_bin()`).

Warning: Removed 2 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



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

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(patients_tble)
```

```
# A tibble: 6 x 6
  subject_id gender anchor_age anchor_year anchor_year_group dod
       <dbl> <chr>       <dbl>       <dbl> <chr>             <date>
1   10000032 F              52        2180 2014 - 2016       2180-09-09
2   10000048 F              23        2126 2008 - 2010       NA
3   10000058 F              33        2168 2020 - 2022       NA
4   10000068 F              19        2160 2008 - 2010       NA
5   10000084 M              72        2160 2017 - 2019       2161-02-13
6   10000102 F              27        2136 2008 - 2010       NA
```

Q4.2 Summary and visualization

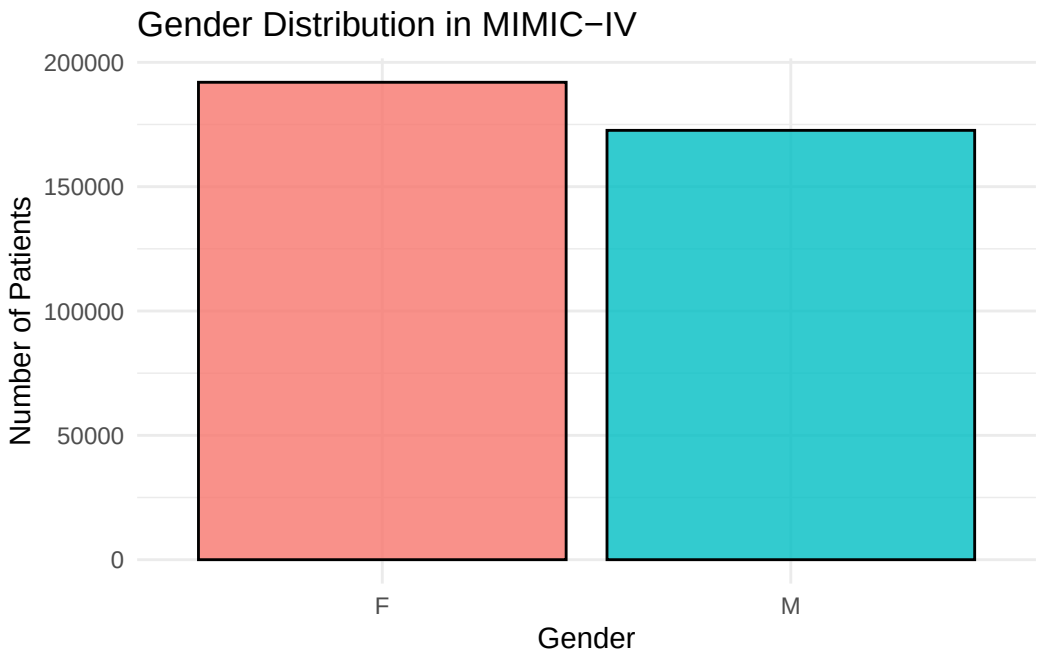
Summarize variables `gender` and `anchor_age` by graphics, and explain any patterns you see.

Solution Slightly more female patients than male, this may be due to longer life expectancy for women. The age distribution is right-skewed, meaning more young and middle-aged patients. The largest group is 20-30 years old, but there is still a significant number of elderly patients.

```r
# Gender
gender_counts <- patients_tble |>
  group_by(gender) |>
  summarise(count = n(), .groups = "drop")
print(gender_counts)
```

```
# A tibble: 2 x 2
  gender  count
  <chr>   <int>
1 F      191984
2 M      172643
```

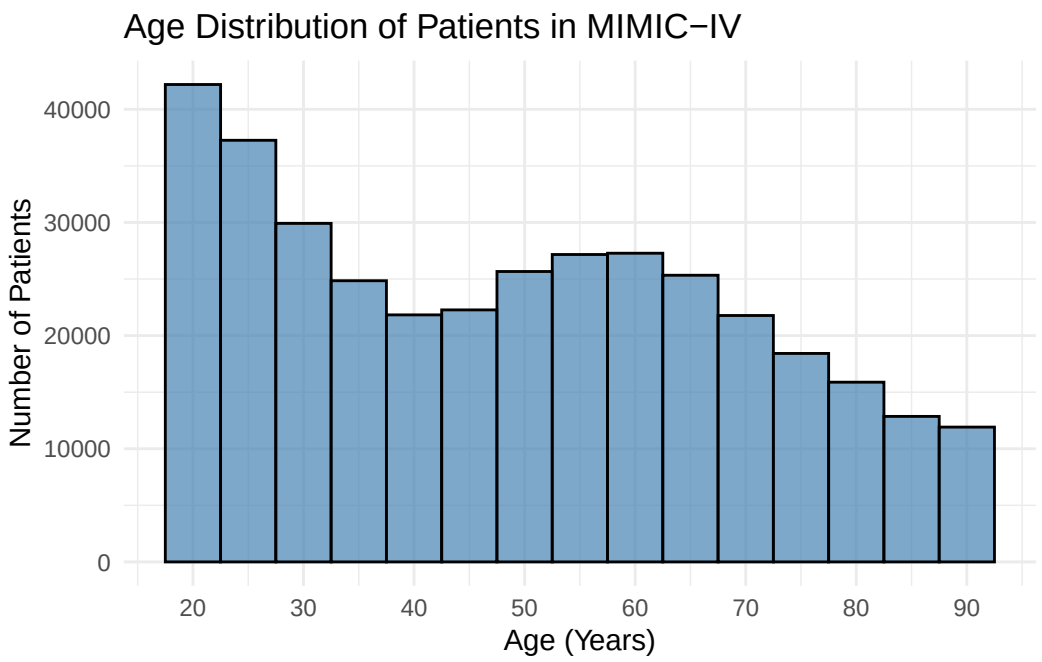
```r
ggplot(gender_counts, aes(x = gender, y = count, fill = gender)) +
  geom_bar(stat = "identity", color = "black", alpha = 0.8) +
  labs(
    title = "Gender Distribution in MIMIC-IV",
    x = "Gender",
    y = "Number of Patients"
  ) +
  theme_minimal() +
  theme(legend.position = "none")
```



```r
# Anchor Age
summary(patients_tble$anchor_age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   29.00   48.00   48.88   65.00   91.00
```

```r
ggplot(patients_tble, aes(x = anchor_age)) +
  geom_histogram(binwidth = 5, fill = "steelblue",
                 color = "black", alpha = 0.7) +
  scale_x_continuous(breaks = seq(0, 100, by = 10)) +
  labs(
    title = "Age Distribution of Patients in MIMIC-IV",
    x = "Age (Years)",
    y = "Number of Patients"
  ) +
  theme_minimal()
```



Q5. Lab results

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

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

```
labevent_id,subject_id,hadm_id,specimen_id,itemid,order_provider_id,charttime,storetime,valu
1,10000032,,2704548,50931,P69FQC,2180-03-23 11:51:00,2180-03-23 15:56:00,___,95,mg/dL,70,100
2,10000032,,36092842,51071,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,
3,10000032,,36092842,51074,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,
4,10000032,,36092842,51075,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,"
5,10000032,,36092842,51079,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,
6,10000032,,36092842,51087,P69FQC,2180-03-23 11:51:00,,,,,,,ROUTINE,RANDOM.
7,10000032,,36092842,51089,P69FQC,2180-03-23 11:51:00,2180-03-23 16:15:00,,,,,,ROUTINE,PRESU
8,10000032,,36092842,51090,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,MI
9,10000032,,36092842,51092,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,ROUTINE,"C
```

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

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

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

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

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

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

Solution

```
icustays_tble <- read_csv("~/mimic/icu/icustays.csv.gz") |>
  select(subject_id, stay_id, intime) |>
  print(width = Inf)
```

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.

# A tibble: 94,458 x 3
   subject_id  stay_id intime
        <dbl>    <dbl> <dttm>
 1   10000032 39553978 2180-07-23 14:00:00
 2   10000690 37081114 2150-11-02 19:37:00
 3   10000980 39765666 2189-06-27 08:42:00
 4   10001217 37067082 2157-11-20 19:18:02
 5   10001217 34592300 2157-12-19 15:42:24
 6   10001725 31205490 2110-04-11 15:52:22
```

```
 7   10001843 39698942 2134-12-05 18:50:03
 8   10001884 37510196 2131-01-11 04:20:05
 9   10002013 39060235 2160-05-18 10:00:53
10   10002114 34672098 2162-02-17 23:30:00
# i 94,448 more rows
```

```
head(icustays_tble)
```

```
# A tibble: 6 x 3
  subject_id  stay_id intime
       <dbl>    <dbl> <dttm>
1   10000032 39553978 2180-07-23 14:00:00
2   10000690 37081114 2150-11-02 19:37:00
3   10000980 39765666 2189-06-27 08:42:00
4   10001217 37067082 2157-11-20 19:18:02
5   10001217 34592300 2157-12-19 15:42:24
6   10001725 31205490 2110-04-11 15:52:22
```

```
labevents_tble <- open_dataset(labevents_parquet, format = "parquet") |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in%
           c(50912, 50971, 50983, 50902, 50882, 51221, 51301, 50931)) |>
  collect()

labevents_tble <- labevents_tble |>
  inner_join(icustays_tble, by = "subject_id","stay_id") |>
  filter(storetime < intime) |>
  group_by(subject_id, stay_id, itemid) |>
  arrange(storetime, .by_group = TRUE) |>
  slice(n()) |>
  ungroup() |>
  select(-c(storetime, intime))
```

```
Warning in inner_join(labevents_tble, icustays_tble, by = "subject_id", : Detected an unexpe
i Row 3958 of `x` matches multiple rows in `y`.
i Row 1 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.
```

```r
labevents_tble_final <- labevents_tble |>
  spread(key = itemid, value = valuenum) |>
  rename(creatinine ="50912",
         potassium ="50971",
         sodium ="50983",
         chloride ="50902",
         bicarbonate ="50882",
         hematocrit ="51221",
         wbc ="51301",
         glucose ="50931") |>
  print(width = Inf)
```

```
# A tibble: 88,086 x 10
   subject_id  stay_id bicarbonate chloride creatinine glucose potassium sodium
        <dbl>    <dbl>       <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
   hematocrit   wbc
        <dbl> <dbl>
 1       41.1   6.9
 2       36.1   7.1
 3       27.3   5.3
 4       37.4   5.4
 5       38.1  15.7
 6       NA    NA
 7       31.4  10.4
 8       39.7  12.2
 9       34.9   7.2
10       34.3  16.8
# i 88,076 more rows
```

Q6. Vitals from charted events

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

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

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

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

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

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

We are interested in the vitals for ICU patients: heart rate (220045), systolic non-invasive blood pressure (220179), diastolic non-invasive blood pressure (220180), body temperature in Fahrenheit (223761), and respiratory rate (220210). Retrieve a subset of `chartevents.csv.gz` only containing these items for the patients in `icustays_tble`. Further restrict to the first

vital measurement within the ICU stay. The final `chartevents_tble` should have one row per ICU stay and columns for each vital measurement.

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

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

Solution

```
chartevents_parquet <- "chartevents_parquet"
chartevents_tble <- open_dataset(chartevents_parquet, format = "parquet") |>
  select(subject_id, itemid, stay_id, storetime, valuenum) |>
  filter(itemid %in% c(220045, 220179, 220180, 223761, 220210)) |>
  collect()

chartevents_tble <- chartevents_tble |>
  inner_join(icustays_tble, by = c("subject_id", "stay_id")) |>
  filter(storetime > intime) |>
  group_by(subject_id, stay_id, itemid) |>
  arrange(storetime, .by_group = TRUE) |>
  slice(1) |>
  ungroup() |>
  select(-c(storetime, intime))

chartevents_tble <- chartevents_tble |>
  spread(key = itemid, value = valuenum) |>
  rename(heart_rate = "220045",
         non_invasive_blood_pressure_systolic = "220179",
         non_invasive_blood_pressure_diastolic = "220180",
         respiratory_rate = "220210",
         temperature_fahrenheit = "223761") |>
  print(width = Inf)
```

```
# A tibble: 94,438 x 7
```

```
   subject_id  stay_id heart_rate non_invasive_blood_pressure_systolic
        <dbl>    <dbl>      <dbl>                                <dbl>
 1   10000032 39553978         91                                   84
 2   10000690 37081114         79                                  107
 3   10000980 39765666         77                                  150
 4   10001217 34592300         96                                  167
 5   10001217 37067082         86                                  151
 6   10001725 31205490         86                                   73
 7   10001843 39698942        118                                  112
 8   10001884 37510196         38                                  180
 9   10002013 39060235         80                                  104
10   10002114 34672098        111                                  112
   non_invasive_blood_pressure_diastolic respiratory_rate temperature_fahrenheit
                                   <dbl>            <dbl>                  <dbl>
 1                                    48               24                   98.7
 2                                    63               23                   97.7
 3                                    77               23                   98
 4                                    95               11                   97.6
 5                                    90               18                   98.5
 6                                    56               19                   97.7
 7                                    71               17                   97.9
 8                                    12               16                   98.1
 9                                    70               14                   97.2
10                                    80               20                   97.9
# i 94,428 more rows
```

Rewrite After Lec@2/20/2025

```
# Rewrite After Lec@2/20/2025
d_items_tble <- read_csv("~/mimic/icu/d_items.csv.gz") |>
  select(itemid, label, abbreviation) |>
  mutate(itemid = as.character(itemid))
```

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Load chartevents from Parquet and process
chartevents_tble <- open_dataset(chartevents_parquet, format = "parquet") |>
  to_duckdb() |>
  select(subject_id, itemid, stay_id, storetime, valuenum) |>
  filter(itemid %in% c(220045, 220179, 220180, 223761, 220210)) |>
  left_join(icustays_tble, by = c("subject_id", "stay_id"), copy = TRUE) |>
  filter(storetime > intime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_max(order_by = storetime, n = 1) |>
  select(-c(storetime, intime)) |>
  ungroup() |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  collect() |>
  arrange(subject_id, stay_id)

# Rename columns based on d_items_tble
chartevents_tble <- chartevents_tble |>
  rename_with(~ str_to_lower(d_items_tble$label
                             [match(.x, d_items_tble$itemid)]),
              .cols = intersect(names(chartevents_tble), d_items_tble$itemid))

col_order <- c("subject_id", "stay_id",
               setdiff(names(chartevents_tble), c("subject_id", "stay_id")))
chartevents_tble <- chartevents_tble |>
relocate(all_of(col_order))

# Print final table
print(chartevents_tble, width = Inf)
```

```
# A tibble: 94,438 x 7
   subject_id  stay_id `temperature fahrenheit`
        <dbl>    <dbl>                    <dbl>
 1   10000032 39553978                     99.5
 2   10000690 37081114                     98
 3   10000980 39765666                     98.7
 4   10001217 34592300                     98.3
 5   10001217 37067082                     99.1
 6   10001725 31205490                     98.4
 7   10001843 39698942                     97.5
 8   10001884 37510196                     99.1
 9   10002013 39060235                     97.8
10   10002114 34672098                     98.2
```

```
   `non invasive blood pressure diastolic`
                                     <dbl>
 1                                      59
 2                                      58
 3                                      69
 4                                      78
 5                                      86
 6                                      58
 7                                      65
 8                                      48
 9                                      60
10                                      79
   `non invasive blood pressure systolic` `heart rate` `respiratory rate`
                                    <dbl>        <dbl>               <dbl>
 1                                     85           94                  20
 2                                     93           90                  26
 3                                    131           69                  21
 4                                    107           80                  22
 5                                    144           93                  17
 6                                     91           73                  23
 7                                     99          136                  27
 8                                     86           74                  14
 9                                    106           94                  14
10                                    147           87                  27
# i 94,428 more rows
```

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 x 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

```
icustays_tble_q7 <-  read_csv("~/mimic/icu/icustays.csv.gz")
```

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

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(icustays_tble_q7)
```

```
# A tibble: 6 x 8
  subject_id  hadm_id  stay_id first_careunit  last_careunit intime
       <dbl>    <dbl>    <dbl> <chr>           <chr>         <dttm>
1   10000032 29079034 39553978 Medical Intens~ Medical Inte~ 2180-07-23 14:00:00
2   10000690 25860671 37081114 Medical Intens~ Medical Inte~ 2150-11-02 19:37:00
3   10000980 26913865 39765666 Medical Intens~ Medical Inte~ 2189-06-27 08:42:00
4   10001217 24597018 37067082 Surgical Inten~ Surgical Int~ 2157-11-20 19:18:02
5   10001217 27703517 34592300 Surgical Inten~ Surgical Int~ 2157-12-19 15:42:24
6   10001725 25563031 31205490 Medical/Surgic~ Medical/Surg~ 2110-04-11 15:52:22
# i 2 more variables: outtime <dttm>, los <dbl>
```

```
head(admissions_tble)
```

```
# A tibble: 6 x 19
  subject_id hadm_id admittime           dischtime           deathtime
```

```
       <dbl>   <dbl> <dttm>              <dttm>              <dttm>
1   10000032  2.26e7 2180-05-06 22:23:00 2180-05-07 17:15:00 NA
2   10000032  2.28e7 2180-06-26 18:27:00 2180-06-27 18:49:00 NA
3   10000032  2.57e7 2180-08-05 23:44:00 2180-08-07 17:50:00 NA
4   10000032  2.91e7 2180-07-23 12:35:00 2180-07-25 17:55:00 NA
5   10000068  2.50e7 2160-03-03 23:16:00 2160-03-04 06:26:00 NA
6   10000084  2.31e7 2160-11-21 01:56:00 2160-11-25 14:52:00 NA
# i 14 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>, admit_hour <int>,
#   admit_minute <int>, length_of_stay <dbl>
```

```
head(patients_tble)
```

```
# A tibble: 6 x 6
  subject_id gender anchor_age anchor_year anchor_year_group dod
       <dbl> <chr>       <dbl>       <dbl> <chr>              <date>
1   10000032 F              52        2180 2014 - 2016       2180-09-09
2   10000048 F              23        2126 2008 - 2010       NA
3   10000058 F              33        2168 2020 - 2022       NA
4   10000068 F              19        2160 2008 - 2010       NA
5   10000084 M              72        2160 2017 - 2019       2161-02-13
6   10000102 F              27        2136 2008 - 2010       NA
```

```
head(labevents_tble_final)
```

```
# A tibble: 6 x 10
  subject_id  stay_id bicarbonate chloride creatinine glucose potassium sodium
       <dbl>    <dbl>       <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
# i 2 more variables: hematocrit <dbl>, wbc <dbl>
```

```
head(chartevents_tble)
```

```
# A tibble: 6 x 7
  subject_id  stay_id `temperature fahrenheit` non invasive blood pressure dia~1
       <dbl>    <dbl>                    <dbl>                              <dbl>
1   10000032 39553978                     99.5                                 59
2   10000690 37081114                     98                                   58
3   10000980 39765666                     98.7                                 69
4   10001217 34592300                     98.3                                 78
5   10001217 37067082                     99.1                                 86
6   10001725 31205490                     98.4                                 58
# i abbreviated name: 1: `non invasive blood pressure diastolic`
# i 3 more variables: `non invasive blood pressure systolic` <dbl>,
#   `heart rate` <dbl>, `respiratory rate` <dbl>
```

```r
labevents_tble_q7 <- labevents_tble_final |>
  group_by(stay_id) |>
  slice(1) |>
  ungroup()

chartevents_tble <- chartevents_tble |>
  group_by(stay_id) |>
  slice(1) |>
  ungroup()

mimic_icu_cohort <- icustays_tble_q7 |>
  left_join(admissions_tble, by = c("subject_id", "hadm_id")) |>
  left_join(patients_tble, by = "subject_id") |>
  mutate(intime_age = anchor_age + year(intime) - anchor_year) |>
  filter(intime_age >= 18) |>
  left_join(labevents_tble_q7, by = "stay_id") |>
  left_join(chartevents_tble, by = "stay_id") |>
  distinct() |>
  print(width = Inf)
```

```
# A tibble: 94,458 x 46
   subject_id.x  hadm_id  stay_id
          <dbl>    <dbl>    <dbl>
 1     10000032 29079034 39553978
 2     10000690 25860671 37081114
 3     10000980 26913865 39765666
 4     10001217 24597018 37067082
 5     10001217 27703517 34592300
 6     10001725 25563031 31205490
```

```
 7     10001843 26133978 39698942
 8     10001884 26184834 37510196
 9     10002013 23581541 39060235
10     10002114 27793700 34672098
   first_careunit
   <chr>
 1 Medical Intensive Care Unit (MICU)
 2 Medical Intensive Care Unit (MICU)
 3 Medical Intensive Care Unit (MICU)
 4 Surgical Intensive Care Unit (SICU)
 5 Surgical Intensive Care Unit (SICU)
 6 Medical/Surgical Intensive Care Unit (MICU/SICU)
 7 Medical/Surgical Intensive Care Unit (MICU/SICU)
 8 Medical Intensive Care Unit (MICU)
 9 Cardiac Vascular Intensive Care Unit (CVICU)
10 Coronary Care Unit (CCU)
   last_careunit                                     intime
   <chr>                                             <dttm>
 1 Medical Intensive Care Unit (MICU)                2180-07-23 14:00:00
 2 Medical Intensive Care Unit (MICU)                2150-11-02 19:37:00
 3 Medical Intensive Care Unit (MICU)                2189-06-27 08:42:00
 4 Surgical Intensive Care Unit (SICU)               2157-11-20 19:18:02
 5 Surgical Intensive Care Unit (SICU)               2157-12-19 15:42:24
 6 Medical/Surgical Intensive Care Unit (MICU/SICU) 2110-04-11 15:52:22
 7 Medical/Surgical Intensive Care Unit (MICU/SICU) 2134-12-05 18:50:03
 8 Medical Intensive Care Unit (MICU)                2131-01-11 04:20:05
 9 Cardiac Vascular Intensive Care Unit (CVICU)      2160-05-18 10:00:53
10 Coronary Care Unit (CCU)                          2162-02-17 23:30:00
   outtime               los admittime           dischtime
   <dttm>              <dbl> <dttm>              <dttm>
 1 2180-07-23 23:50:47 0.410 2180-07-23 12:35:00 2180-07-25 17:55:00
 2 2150-11-06 17:03:17 3.89  2150-11-02 18:02:00 2150-11-12 13:45:00
 3 2189-06-27 20:38:27 0.498 2189-06-27 07:38:00 2189-07-03 03:00:00
 4 2157-11-21 22:08:00 1.12  2157-11-18 22:56:00 2157-11-25 18:00:00
 5 2157-12-20 14:27:41 0.948 2157-12-18 16:58:00 2157-12-24 14:55:00
 6 2110-04-12 23:59:56 1.34  2110-04-11 15:08:00 2110-04-14 15:00:00
 7 2134-12-06 14:38:26 0.825 2134-12-05 00:10:00 2134-12-06 12:54:00
 8 2131-01-20 08:27:30 9.17  2131-01-07 20:39:00 2131-01-20 05:15:00
 9 2160-05-19 17:33:33 1.31  2160-05-18 07:45:00 2160-05-23 13:30:00
10 2162-02-20 21:16:27 2.91  2162-02-17 22:32:00 2162-03-04 15:16:00
   deathtime           admission_type              admit_provider_id
   <dttm>              <chr>                       <chr>
 1 NA                  EW EMER.                    P06OTX
```

```
 2 NA                  EW EMER.                    P26QQ4
 3 NA                  EW EMER.                    P06OTX
 4 NA                  EW EMER.                    P3610N
 5 NA                  DIRECT EMER.                P276OU
 6 NA                  EW EMER.                    P32W56
 7 2134-12-06 12:54:00 URGENT                      P67ATB
 8 2131-01-20 05:15:00 OBSERVATION ADMIT           P49AFC
 9 NA                  SURGICAL SAME DAY ADMISSION P8286C
10 NA                  OBSERVATION ADMIT           P46834
   admission_location     discharge_location insurance language marital_status
   <chr>                  <chr>              <chr>     <chr>    <chr>
 1 EMERGENCY ROOM         HOME               Medicaid  English  WIDOWED
 2 EMERGENCY ROOM         REHAB              Medicare  English  WIDOWED
 3 EMERGENCY ROOM         HOME HEALTH CARE   Medicare  English  MARRIED
 4 EMERGENCY ROOM         HOME HEALTH CARE   Private   Other    MARRIED
 5 PHYSICIAN REFERRAL     HOME HEALTH CARE   Private   Other    MARRIED
 6 PACU                   HOME               Private   English  MARRIED
 7 TRANSFER FROM HOSPITAL DIED               Medicare  English  SINGLE
 8 EMERGENCY ROOM         DIED               Medicare  English  MARRIED
 9 PHYSICIAN REFERRAL     HOME HEALTH CARE   Medicare  English  SINGLE
10 PHYSICIAN REFERRAL     HOME HEALTH CARE   Medicaid  English  <NA>
   race                   edregtime           edouttime
   <chr>                  <dttm>              <dttm>
 1 WHITE                  2180-07-23 05:54:00 2180-07-23 14:00:00
 2 WHITE                  2150-11-02 11:41:00 2150-11-02 19:37:00
 3 BLACK/AFRICAN AMERICAN 2189-06-27 06:25:00 2189-06-27 08:42:00
 4 WHITE                  2157-11-18 17:38:00 2157-11-19 01:24:00
 5 WHITE                  NA                  NA
 6 WHITE                  NA                  NA
 7 WHITE                  NA                  NA
 8 BLACK/AFRICAN AMERICAN 2131-01-07 13:36:00 2131-01-07 22:13:00
 9 OTHER                  NA                  NA
10 UNKNOWN                2162-02-17 19:35:00 2162-02-17 23:30:00
   hospital_expire_flag admit_hour admit_minute length_of_stay gender anchor_age
                  <dbl>      <int>        <int>          <dbl> <chr>       <dbl>
 1                    0         12           35           2.22 F              52
 2                    0         18            2           9.82 F              86
 3                    0          7           38           5.81 F              73
 4                    0         22           56           6.79 F              55
 5                    0         16           58           5.91 F              55
 6                    0         15            8           2.99 F              46
 7                    1          0           10           1.53 M              73
 8                    1         20           39          12.4  F              68
```

```
 9                   0          7           45           5.24 F              53
10                   0         22           32          14.7  M              56
   anchor_year anchor_year_group dod        intime_age subject_id.y bicarbonate
         <dbl> <chr>             <date>          <dbl>        <dbl>       <dbl>
 1        2180 2014 - 2016       2180-09-09         52     10000032          25
 2        2150 2008 - 2010       2152-01-30         86     10000690          26
 3        2186 2008 - 2010       2193-08-26         76     10000980          21
 4        2157 2011 - 2013       NA                 55     10001217          22
 5        2157 2011 - 2013       NA                 55     10001217          30
 6        2110 2011 - 2013       NA                 46     10001725          NA
 7        2131 2017 - 2019       2134-12-06         76     10001843          28
 8        2122 2008 - 2010       2131-01-20         77     10001884          30
 9        2156 2008 - 2010       NA                 57     10002013          24
10        2162 2020 - 2022       2162-12-11         56     10002114          18
   chloride creatinine glucose potassium sodium hematocrit   wbc subject_id
      <dbl>      <dbl>   <dbl>     <dbl>  <dbl>      <dbl> <dbl>      <dbl>
 1       95        0.7     102       6.7    126       41.1   6.9   10000032
 2      100        1        85       4.8    137       36.1   7.1   10000690
 3      109        2.3      89       3.9    144       27.3   5.3   10000980
 4      108        0.6     112       4.2    142       38.1  15.7   10001217
 5      104        0.5      87       4.1    142       37.4   5.4   10001217
 6       98       NA        NA       4.1    139       NA    NA     10001725
 7       97        1.3     131       3.9    138       31.4  10.4   10001843
 8       88        1.1     141       4.5    130       39.7  12.2   10001884
 9      102        0.9     288       3.5    137       34.9   7.2   10002013
10       NA        3.1      95       6.5    125       34.3  16.8   10002114
   `temperature fahrenheit` `non invasive blood pressure diastolic`
                      <dbl>                                   <dbl>
 1                     99.5                                      59
 2                     98                                        58
 3                     98.7                                      69
 4                     99.1                                      86
 5                     98.3                                      78
 6                     98.4                                      58
 7                     97.5                                      65
 8                     99.1                                      48
 9                     97.8                                      60
10                     98.2                                      79
   `non invasive blood pressure systolic` `heart rate` `respiratory rate`
                                    <dbl>        <dbl>              <dbl>
 1                                     85           94                 20
 2                                     93           90                 26
 3                                    131           69                 21
```

```
 4                                  144           93                  17
 5                                  107           80                  22
 6                                   91           73                  23
 7                                   99          136                  27
 8                                   86           74                  14
 9                                  106           94                  14
10                                  147           87                  27
# i 94,448 more rows
```

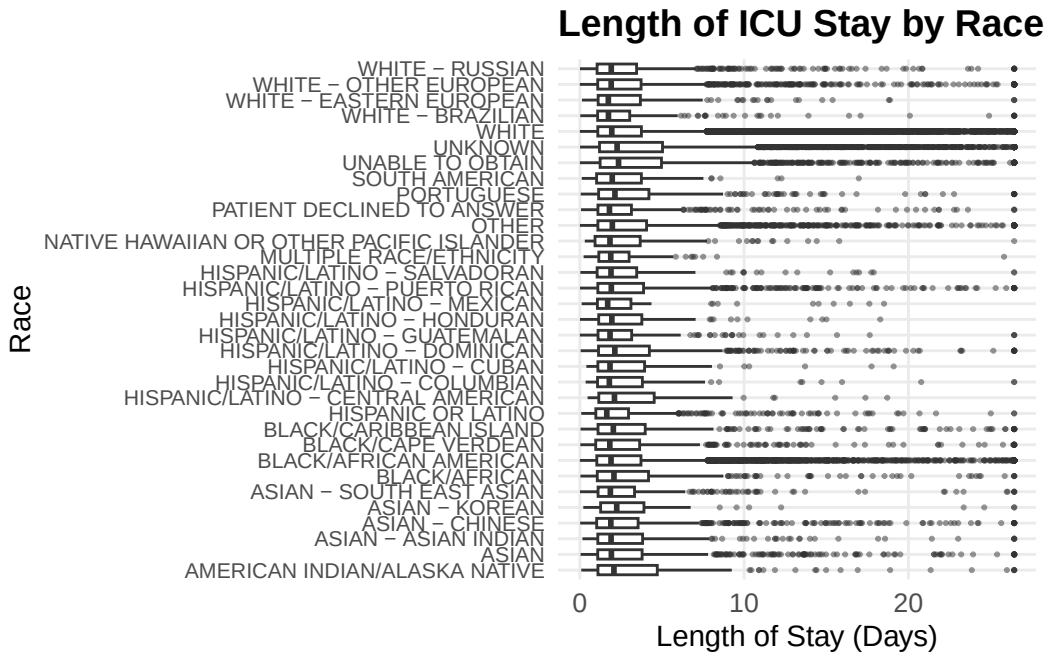
Q8. Exploratory data analysis (EDA)

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) **Solution**

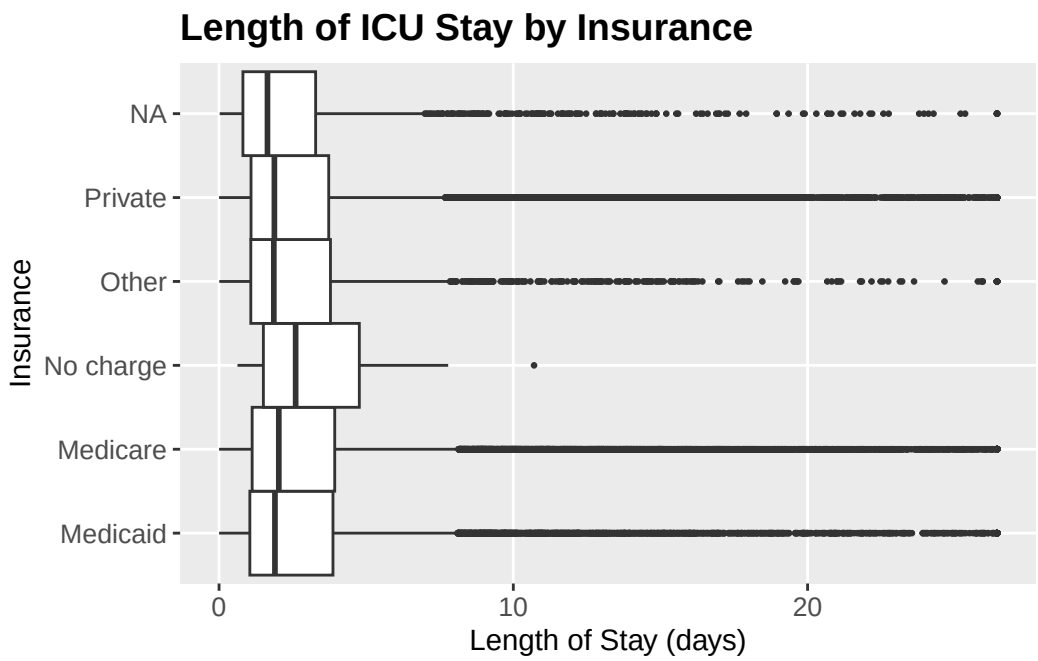
```
mimic_icu_cohort <- mimic_icu_cohort |>
  mutate(race = as.factor(race),
         insurance = as.factor(insurance),
         marital_status = as.factor(marital_status),
         gender = as.factor(gender))
# LOS vs Race
mimic_icu_cohort |>
  ggplot(mapping = aes(x = race, y = los)) +
  geom_boxplot(width = 0.6, outlier.size = 0.4, outlier.alpha = 0.5) +
  coord_flip() +
  labs(
    title = "Length of ICU Stay by Race",
    x = "Race",
    y = "Length of Stay (Days)"
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 8),
    axis.text.x = element_text(size = 10),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.minor = element_blank()
  ) +
  scale_y_continuous(limits = c(0, quantile(mimic_icu_cohort$los, 0.99,
                                            na.rm = TRUE)),
                     oob = scales::squish)
```

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



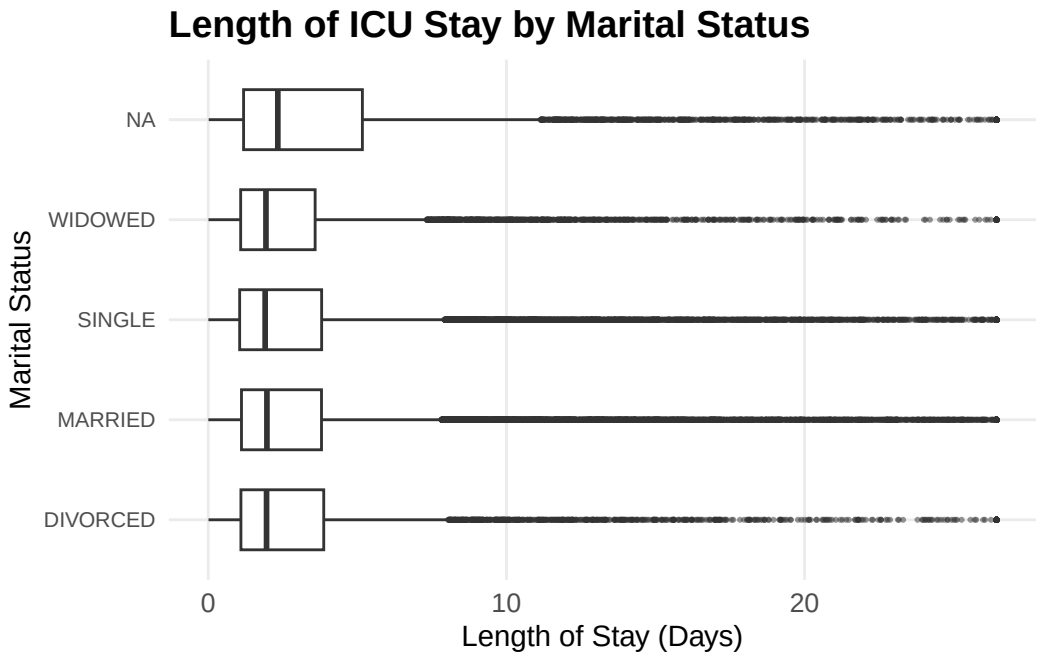

```r
# LOS vs Insurance
mimic_icu_cohort |>
  ggplot(mapping = aes(x = insurance, y = los)) +
  geom_boxplot(width = 1, outlier.size = 0.5,
               position = position_dodge(width = 0.5)) +
  coord_flip() +
  labs(
    title = "Length of ICU Stay by Insurance",
    x = "Insurance",
    y = "Length of Stay (days)"
  ) +
  theme(
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(size = 10),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.minor = element_blank()
  ) +
  scale_y_continuous(limits = c(0, quantile(mimic_icu_cohort$los, 0.99,
                                            na.rm = TRUE)),
                     oob = scales::squish)
```

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



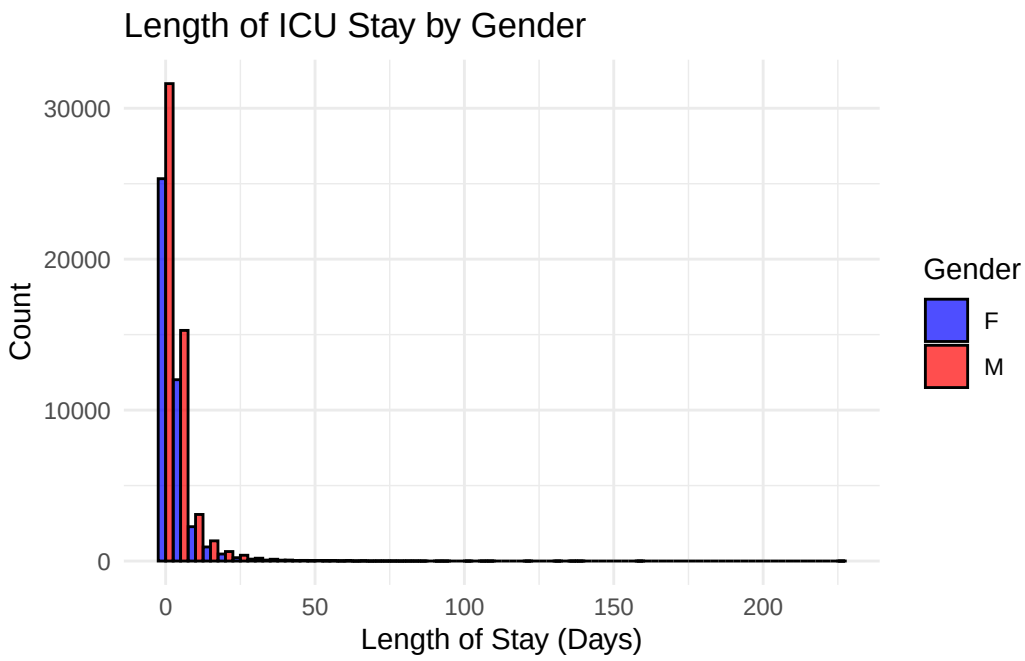

```
# LOS vs Marital Status
mimic_icu_cohort |>
  ggplot(mapping = aes(x = marital_status, y = los)) +
  geom_boxplot(width = 0.6, outlier.size = 0.4, outlier.alpha = 0.5) +
  coord_flip() +
  labs(
    title = "Length of ICU Stay by Marital Status",
    x = "Marital Status",
    y = "Length of Stay (Days)"
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 8),
    axis.text.x = element_text(size = 10),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.minor = element_blank()
  ) +
  scale_y_continuous(limits = c(0, quantile(mimic_icu_cohort$los, 0.99,
                                            na.rm = TRUE)),
                     oob = scales::squish)
```

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




```r
# LOS vs gender
mimic_icu_cohort |>
  ggplot(aes(x = los, fill = gender)) +
  geom_histogram(position = "dodge", binwidth = 5,
                 color = "black", alpha = 0.7) +
  labs(title = "Length of ICU Stay by Gender",
       x = "Length of Stay (Days)",
       y = "Count",
       fill = "Gender") +
  scale_fill_manual(values = c("blue", "red")) +
  theme_minimal()
```

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

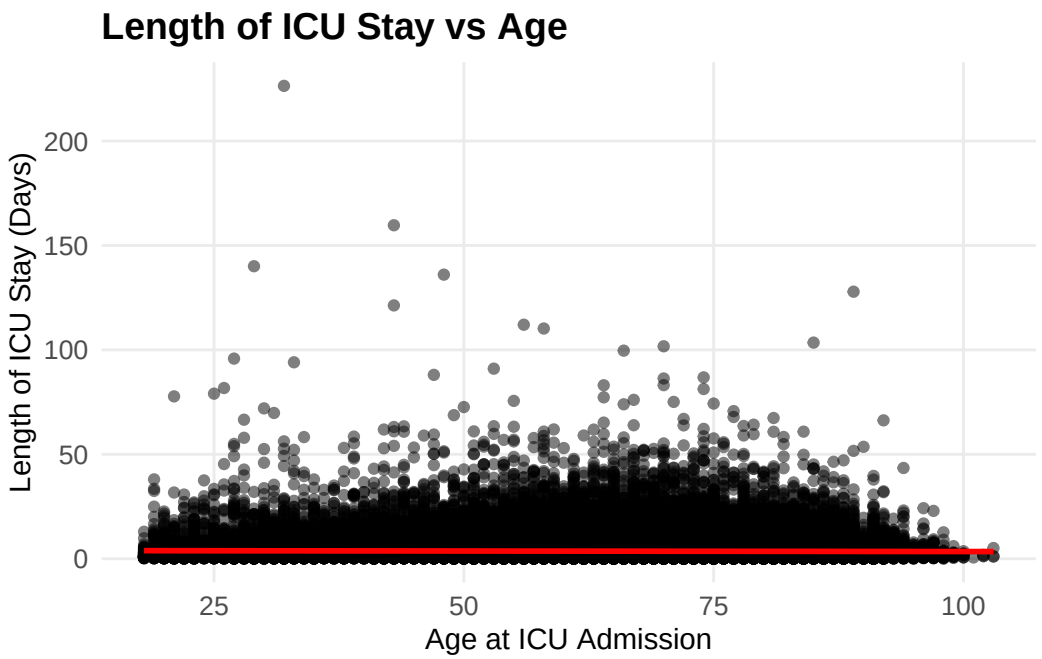



```
# LOS vs Age at Intime
ggplot(mimic_icu_cohort, aes(x = intime_age, y = los)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  labs(title = "Length of ICU Stay vs Age",
       x = "Age at ICU Admission",
       y = "Length of ICU Stay (Days)") +
  theme_minimal()+
  theme(
    axis.text.y = element_text(size = 10),
    axis.text.x = element_text(size = 10),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.minor = element_blank()
  )
```

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

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

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



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

```
library(gridExtra)
```

```
Attaching package: 'gridExtra'

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

    combine
```

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

for (var in lab_vars) {
  p <- ggplot(mimic_icu_cohort, aes_string(x = var, y = "los")) +
    geom_point(alpha = 0.3, color = "blue") +
    geom_smooth(method = "lm", color = "red", se = FALSE) +
    labs(x = var, y = "Length of ICU Stay (LOS)",
         title = paste("ICU Stay vs", var)) +
    theme_minimal()
```

```
  plots[[var]] <- p
}
```

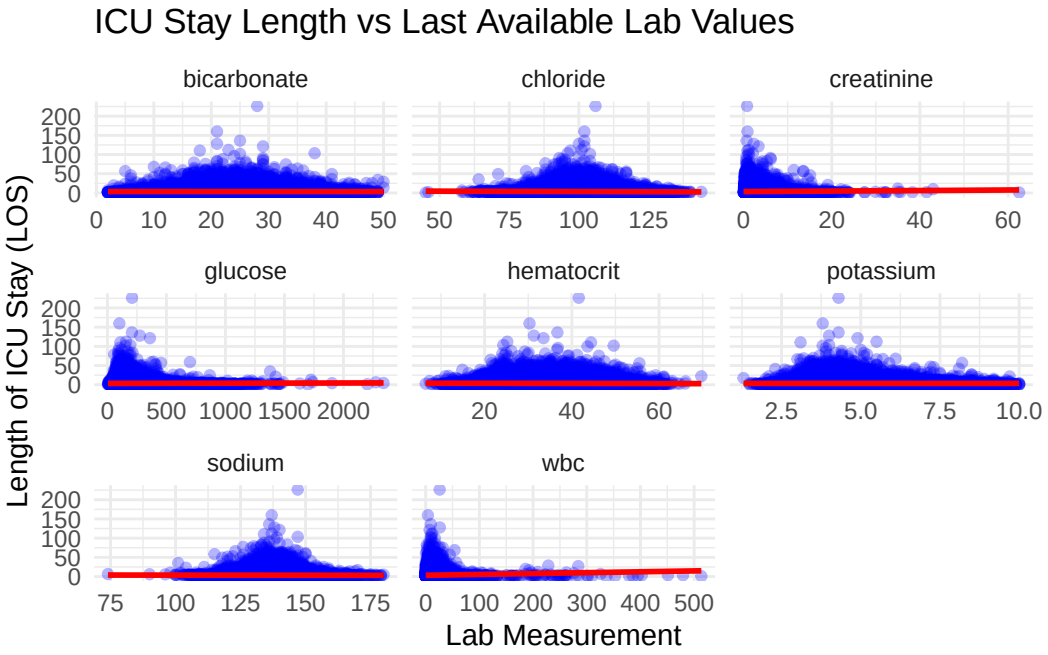
```
Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
i Please use tidy evaluation idioms with `aes()`.
i See also `vignette("ggplot2-in-packages")` for more information.
```

```
mimic_icu_cohort |>
  pivot_longer(cols = all_of(lab_vars),
               names_to = "Lab Test", values_to = "Value") |>
  ggplot(aes(x = Value, y = los)) +
  geom_point(alpha = 0.3, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  facet_wrap(~`Lab Test`, scales = "free_x") +
  labs(title = "ICU Stay Length vs Last Available Lab Values",
       x = "Lab Measurement", y = "Length of ICU Stay (LOS)") +
  theme_minimal()
```

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

```
Warning: Removed 79011 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 79011 rows containing missing values or values outside the scale range
(`geom_point()`).
```



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

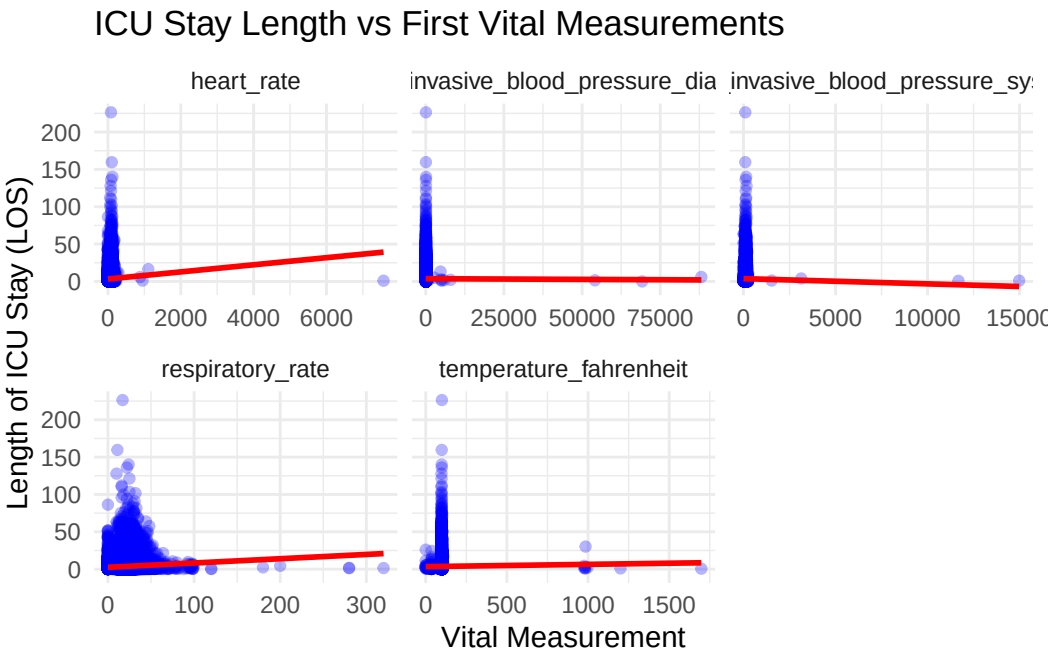
```
mimic_icu_cohort <- mimic_icu_cohort |>
  rename(
    heart_rate = `heart rate`,
    non_invasive_blood_pressure_systolic =
      `non invasive blood pressure systolic`,
    non_invasive_blood_pressure_diastolic =
      `non invasive blood pressure diastolic`,
    respiratory_rate = `respiratory rate`,
    temperature_fahrenheit = `temperature fahrenheit`
  )

vital_vars <- c("heart_rate",
                "non_invasive_blood_pressure_systolic",
                "non_invasive_blood_pressure_diastolic",
                "respiratory_rate",
                "temperature_fahrenheit")

plots <- list()
for (var in vital_vars) {
  p <- ggplot(mimic_icu_cohort, aes_string(x = var, y = "los")) +
    geom_point(alpha = 0.3, color = "blue") +
```

```
    geom_smooth(method = "lm", color = "red", se = FALSE) +
    labs(x = var, y = "Length of ICU Stay (LOS)",
         title = paste("ICU Stay vs", var)) +
    theme_minimal()

  plots[[var]] <- p
}

mimic_icu_cohort |>
  pivot_longer(cols = all_of(vital_vars),
               names_to = "Vital Sign", values_to = "Value") |>
  ggplot(aes(x = Value, y = los)) +
  geom_point(alpha = 0.3, color = "blue") +
  geom_smooth(method = "lm", color = "red", se = FALSE) +
  facet_wrap(~ `Vital Sign`, scales = "free_x") +
  labs(title = "ICU Stay Length vs First Vital Measurements",
       x = "Vital Measurement", y = "Length of ICU Stay (LOS)") +
  theme_minimal()
```

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

```
Warning: Removed 4512 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 4512 rows containing missing values or values outside the scale range
(`geom_point()`).
```



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

```
unique(mimic_icu_cohort$first_careunit)
```

```
 [1] "Medical Intensive Care Unit (MICU)"
 [2] "Surgical Intensive Care Unit (SICU)"
 [3] "Medical/Surgical Intensive Care Unit (MICU/SICU)"
 [4] "Cardiac Vascular Intensive Care Unit (CVICU)"
 [5] "Coronary Care Unit (CCU)"
 [6] "Neuro Intermediate"
 [7] "Trauma SICU (TSICU)"
 [8] "Neuro Stepdown"
 [9] "Neuro Surgical Intensive Care Unit (Neuro SICU)"
[10] "Surgery/Vascular/Intermediate"
[11] "Intensive Care Unit (ICU)"
[12] "PACU"
[13] "Medicine"
[14] "Surgery/Trauma"
[15] "Medicine/Cardiology Intermediate"
[16] "Med/Surg"
[17] "Neurology"
```

```r
mimic_icu_cohort |>
  ggplot(aes(x = first_careunit, y = los)) +
  geom_boxplot(width = 0.8, outlier.size = 0.4, outlier.alpha = 0.5) +
  coord_flip() +
  labs(
    title = "Length of ICU Stay by First Care Unit",
    x = "First Care Unit",
    y = "Length of Stay (Days)"
  ) +
  theme_minimal() +
  theme(
    axis.text.y = element_text(size = 8),
    axis.text.x = element_text(size = 8),
    plot.title = element_text(size = 14, face = "bold"),
    panel.grid.minor = element_blank()
  )
```

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