

Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

AUTHOR

Julie Lee, 806409381

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: Etc/UTC

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_1.0.3	

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

```
library(arrow)
```

Attaching package: 'arrow'

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

timestamp

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

Attaching package: 'pryr'

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

where

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

Loading required package: R.oo

Loading required package: R.methodsS3

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

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

Attaching package: 'R.oo'

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

throw

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

getClasses, getMethods

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

attach, detach, load, save

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

Attaching package: 'R.utils'

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

timestamp

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

timestamp

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

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

```
library(tidyverse)
```

— Attaching core tidyverse packages — tidyverse 2.0.0 —

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

— Conflicts — tidyverse_conflicts() —

```
* purrr::compose()      masks pryr::compose()
* lubridate::duration() masks arrow::duration()
* tidyr::extract()      masks R.utils::extract()
* dplyr::filter()       masks stats::filter()
* dplyr::lag()           masks stats::lag()
* purrr::partial()      masks pryr::partial()
* dplyr::where()         masks pryr::where(), gtsummary::where()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

```
library(duckdb)
```

Loading required package: DBI

Display your machine memory.

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

Totalram: 62.794 GiB

Freeram: 61.471 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.

```
system("ls -ld ./labevents_pq")

patient_data <- 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.

```
hospital_admissions <- 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

dtm (5): admittime, disctime, 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.

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

Rows: 2413581 Columns: 7

— Column specification —

Delimiter: ","

chr (2): eventtype, careunit

dbl (3): subject_id, hadm_id, transfer_id

dtm (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.

```
lab_results <- read_parquet("./labevents_pq/part-0.parquet")
medical_procedures <- read_csv("~/mimic/hosp/procedures_icd.csv.gz")
```

Rows: 859655 Columns: 6

— Column specification —————

Delimiter: ","

chr (1): icd_code

dbl (4): subject_id, hadm_id, seq_num, icd_version

date (1): chartdate

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

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

Rows: 6364488 Columns: 5

— Column specification —————

Delimiter: ","

chr (1): icd_code

dbl (4): subject_id, hadm_id, seq_num, icd_version

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

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

Rows: 86423 Columns: 3

— Column specification —————

Delimiter: ","

chr (2): icd_code, long_title

dbl (1): icd_version

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

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

Rows: 112107 Columns: 3

— Column specification —————

Delimiter: ","

chr (2): icd_code, long_title

dbl (1): icd_version

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

```
icu_stays <- 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

dtm (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.

```
selected_patient <- 10063848
patient_record <- patient_data %>%
  filter(subject_id == !!selected_patient)
admission_details <- hospital_admissions %>%
  filter(subject_id == !!selected_patient)
transfer_details <- patient_transfers %>%
  filter(subject_id == !!selected_patient)
lab_details <- lab_results %>%
  filter(subject_id == !!selected_patient)
procedure_details <- medical_procedures %>%
  filter(subject_id == !!selected_patient)
diagnosis_details <- patient_diagnoses %>%
  filter(subject_id == !!selected_patient)
icu_details <- icu_stays %>%
  filter(subject_id == !!selected_patient)
```

```
diagnosis_details <- diagnosis_details %>%
  mutate(icd_code = str_pad(icd_code, width = 5, side = "left", pad = "0")) %>%
  left_join(diagnosis_codes, by = c("icd_code", "icd_version"))

diagnosis_column <- grep("long_title", names(diagnosis_details), value = TRUE)

if ("long_title" %in% diagnosis_column) {
  diagnosis_details <- rename(diagnosis_details,
                             diagnosis_name = long_title)
} else if ("long_title.x" %in% diagnosis_column) {
  diagnosis_details <- rename(diagnosis_details,
                             diagnosis_name = long_title.x)
} else if ("long_title.y" %in% diagnosis_column) {
  diagnosis_details <- rename(diagnosis_details,
                             diagnosis_name = long_title.y)
} else {
  stop("No appropriate long_title column found in diagnosis_details")
}

top_conditions <- diagnosis_details %>%
  filter(!is.na(diagnosis_name)) %>%
```

```

count(diagnosis_name, sort = TRUE) %>%
slice_head(n = 3) %>%
pull(diagnosis_name)

top_conditions_text <- if (length(top_conditions) > 0) {
  paste(top_conditions, collapse = "\n")
} else { "" }

patient_summary_text <- str_squish(paste0(
  "Patient ", patient_record, ", ",
  ifelse(is.na(patient_record$gender), "",
    paste0(patient_record$gender, ", ")),
  ifelse(is.na(patient_record$anchor_age), "",
    paste0(patient_record$anchor_age, " years old, ")),
  ifelse(is.na(admission_details$race), "", admission_details$race)
))

transfer_details <- transfer_details %>%
  mutate(intime = as.POSIXct(intime, format = "%Y-%m-%d %H:%M:%S"),
    outtime = as.POSIXct(outtime, format = "%Y-%m-%d %H:%M:%S")) %>%
  filter(!is.na(outtime))

lab_details <- lab_details %>%
  mutate(chartdate = as.POSIXct(charttime, format = "%Y-%m-%d %H:%M:%S"))

procedure_details <- procedure_details %>%
  mutate(chartdate = as.POSIXct(chartdate, format = "%Y-%m-%d %H:%M:%S")) %>%
  left_join(procedure_codes, by = c("icd_code", "icd_version"))

procedure_name_column <- grep("long_title", names(procedure_details),
  value = TRUE)

if (length(procedure_name_column) > 1) {
  procedure_details <- procedure_details %>%
    select(-all_of(procedure_name_column[-1])) %>%
    rename(procedure_name = first(procedure_name_column))
} else {
  procedure_details <- rename(procedure_details,
    procedure_name = procedure_name_column)
}

```

Warning: Using an external vector in selections was deprecated in tidysselect 1.1.0.

i Please use `all_of()` or `any_of()` instead.

Was:

```
data %>% select(procedure_name_column)
```

Now:

```
data %>% select(all_of(procedure_name_column))
```

See <<https://tidysselect.r-lib.org/reference/faq-external-vector.html>>.

```

procedure_details <- procedure_details %>%
  filter(!is.na(procedure_name))

distinct_care_units <- unique(transfer_details$careunit)
distinct_procedures <- unique(procedure_details$procedure_name)

care_unit_palette <- setNames(
  RColorBrewer::brewer.pal(n = min(length(distinct_care_units), 9),
    name = "Set1"),
  distinct_care_units
)

transfer_details <- transfer_details %>%
  mutate(careunit = factor(careunit, levels = distinct_care_units))

procedure_shapes <- setNames(
  seq(15, 15 + length(distinct_procedures) - 1),
  distinct_procedures
)

procedure_details <- procedure_details %>%
  mutate(procedure_name = factor(procedure_name, levels = distinct_procedures))

patient_timeline <- ggplot() +
  geom_segment(data = transfer_details,
    aes(x = intime, xend = outtime, y = "ADT", yend = "ADT",
      color = careunit), linewidth = 3) +
  geom_point(data = lab_details,
    aes(x = chartdate, y = "Lab"), shape = 3, size = 3) +
  geom_point(data = procedure_details,
    aes(x = chartdate, y = "Procedure", shape = procedure_name),
    size = 5) +
  scale_color_manual(values = care_unit_palette) +
  scale_shape_manual(values = procedure_shapes, drop = FALSE) +
  scale_y_discrete(limits = c("Procedure", "Lab", "ADT")) +
  theme_minimal() +
  labs(title = patient_summary_text,
    subtitle = top_conditions_text,
    x = "Calendar Time",
    y = NULL,
    color = "Care Unit",
    shape = "Procedure") +
  guides(
    color = guide_legend(order = 1, title.position = "top"),
    shape = guide_legend(order = 2, title.position = "top")
  ) +
  theme(
    legend.position = "bottom",
    legend.box = "vertical",
    legend.box.just = "center",

```



```

legend.spacing.y = unit(0.5, "cm"),
legend.title = element_text(size = 12, face = "bold"),
legend.text = element_text(size = 10)
)

print(patient_timeline)

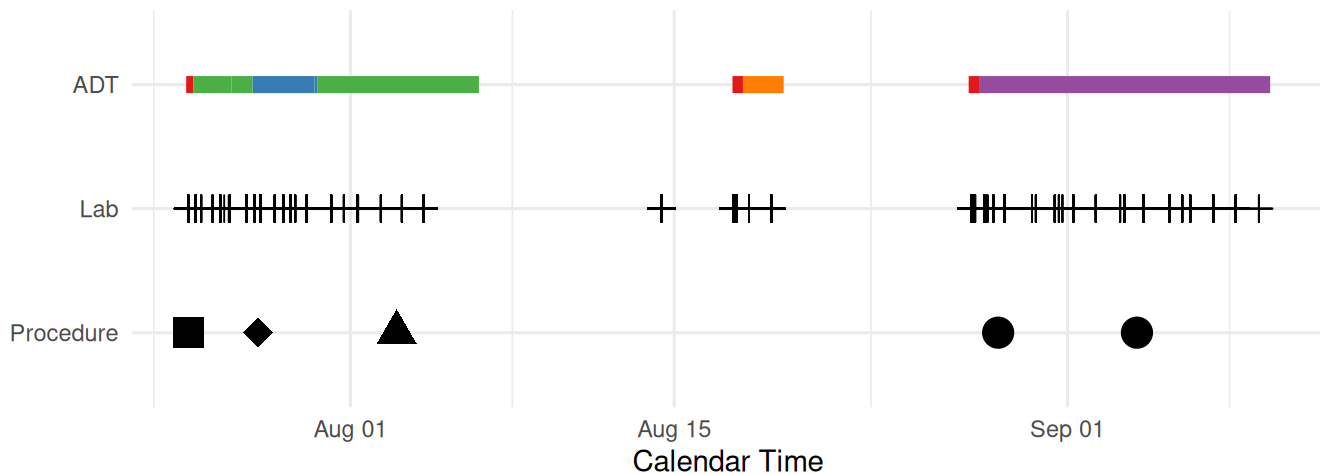
```

Patient 10063848, F, 75 years old, WHITE

Unspecified Escherichia coli [E. coli] as the cause of diseases classified elsewhere

Acute respiratory failure with hypoxia

Adverse effect of sulfonamides, initial encounter



Care Unit

■ Emergency Department
 ■ Surgical Intensive Care Unit (SICU)
 ■ Med/Surg/Trauma
 ■ Medicine
 ■

■ Insertion of Infusion Device into Subclavian Vein
 ◆ Insertion of Infusion Device into Subclavian Vein
 ▲ Insertion of Cardiac Sampling and Pressure, Right Heart, Percutaneous Approach

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 Note that the dataset that I selected is the filtered chart_events parquet file that was created in Homework #2 with the modification of adding in stay_id as well as including the item_id “Diastolic Non-Invasive Blood pressure (2210180) instead of”Mean Non-Invasive Blood Pressure “(2200181).

```

parquet_file <- "./chartevents_pq/part-0.parquet"
start_time <- Sys.time()

```

```

arrow_data <- open_dataset(parquet_file, format = "parquet")
duckdb_conn <- dbConnect(duckdb::duckdb())

invisible(to_duckdb(
  .data = arrow_data,
  con = duckdb_conn,
  table_name = "chartevents",
  auto_disconnect = FALSE
))
subset_chartevents <- tbl(duckdb_conn, "chartevents") %>%
  filter(itemid %in% c(220045, 220179, 220180, 223761, 220210)) %>%
  arrange(subject_id, charttime, itemid,) %>%
  collect()

dbDisconnect(duckdb_conn)

```

```
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

dtm (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.

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

Rows: 4095 Columns: 9

— Column specification —————

Delimiter: ","

chr (6): label, abbreviation, linksto, category, unitname, param_type

dbl (3): itemid, lownormalvalue, highnormalvalue

i Use `spec()` to retrieve the full column specification for this data.

i Specify the column types or set `show_col_types = FALSE` to quiet this message.

```

subject_id_of_interest <- 10063848

subject_stays <- icustays_tble %>%
  filter(subject_id == subject_id_of_interest) %>%
  select(stay_id, intime, outtime) %>%
  collect()

chartevents_filtered <- subset_chartevents %>%
  filter(stay_id %in% subject_stays$stay_id) %>%
  inner_join(subject_stays, by = "stay_id") %>%

```

```

filter(charttime >= intime & charttime <= outtime) %>%
select(stay_id, itemid, charttime, valuenum)

chartevents_with_labels <- chartevents_filtered %>%
  inner_join(d_items %>% select(itemid, abbreviation), by = "itemid") %>%
  mutate(charttime = as_datetime(charttime))

ggplot(chartevents_with_labels,
  aes(x = charttime, y = valuenum, color = abbreviation)) +
  geom_point(size = 1.2) +
  geom_line(size = 0.8) +
  facet_grid(abbreviation ~ stay_id, scales = "free") +
  labs(
    title = paste("Patient", subject_id_of_interest, "ICU stays - Vitals"),
    x = "Time",
    y = "Vital Value"
  ) +
  scale_x_datetime(
    breaks = seq(
      floor_date(min(chartevents_with_labels$charttime, na.rm = TRUE),
        unit = "6 hours"),
      ceiling_date(max(chartevents_with_labels$charttime, na.rm = TRUE),
        unit = "6 hours"),
      by = "6 hours"
    ),
    date_labels = "%b %d %H:%M"
  ) +
  theme_minimal() +
  theme(
    legend.position = "none",
    strip.text = element_text(size = 12, face = "bold", color = "white"),
    strip.background = element_rect(fill = "darkgrey", color = "darkgrey"),
    axis.text.x = element_text(angle = 0, hjust = 0.5),
    panel.grid.major = element_line(size = 0.5, linetype = "dotted",
      color = "gray"),
    panel.grid.minor = element_blank()
  )

```

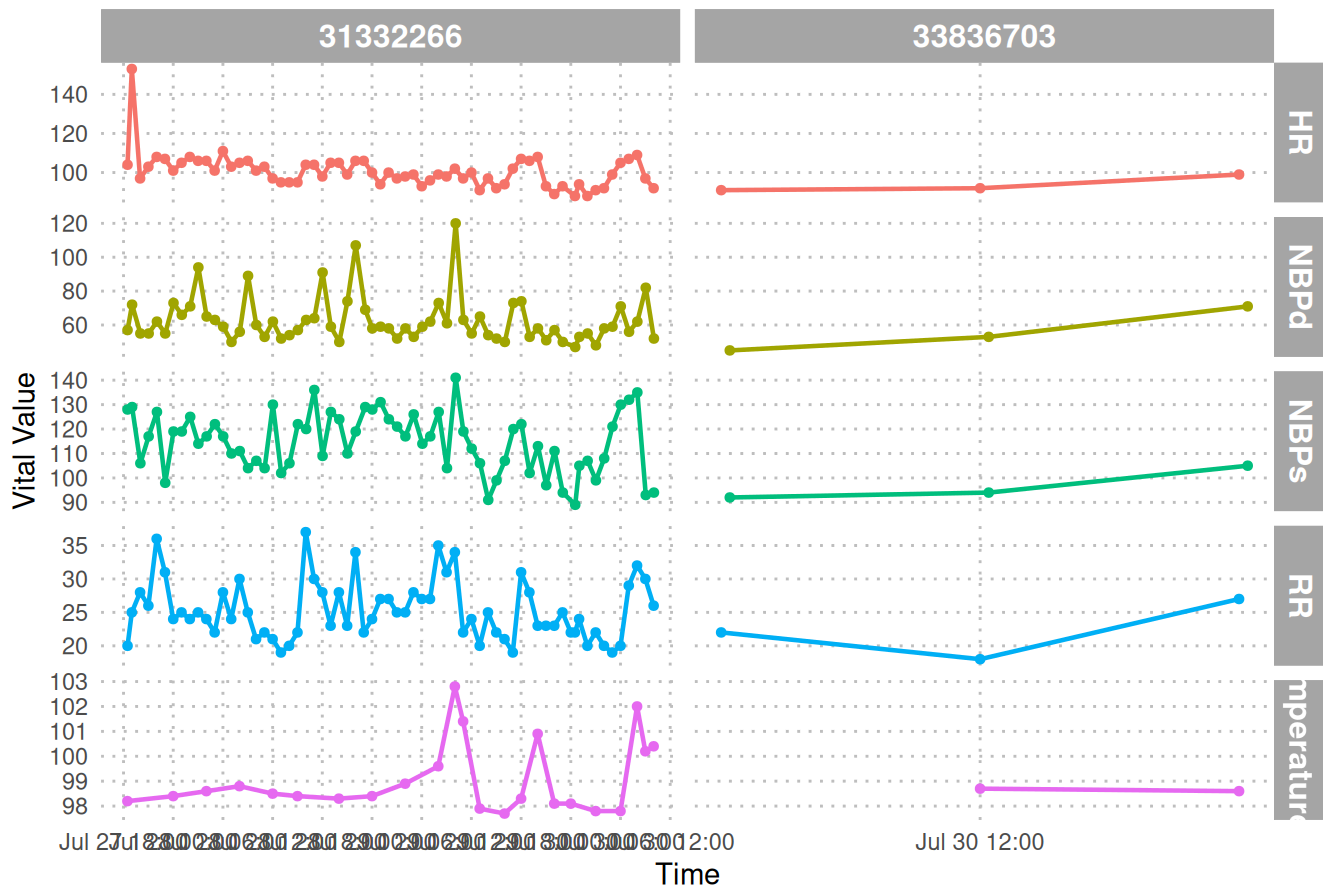
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.

ℹ Please use `linewidth` instead.

Warning: The `size` argument of `element_line()` is deprecated as of ggplot2 3.4.0.

ℹ Please use the `linewidth` argument instead.

Patient 10063848 ICU stays - Vitals



Q2. ICU stays

[icustays.csv.gz](https://mimic.mit.edu/docs/iv/modules/icu/icustays/) (<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`. Done

```
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

dtm (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 × 8

	subject_id <dbl>	hadm_id <dbl>	stay_id <dbl>	first_careunit <chr>	last_careunit <chr>	intime <dtm>
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 <dtm>, 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.

(2.2) Solution

The number of unique `subject_id` there are in the (`icustays_tble`) tibble is 65366 unique subjects.

```
num_unique_subjects <- icustays_tble %>%
  distinct(subject_id) %>%
  nrow()
```

```
print(paste(num_unique_subjects))
```

```
[1] "65366"
```

A subject_id can have multiple ICU stays. We first obtained a tibble with the number of ICU stays per subject_id. Next we filtered out the rows where the subjects only had 1 ICU stay and counted the number of remaining rows. The number of patients with multiple ICU stays is 16242.

```
icu_stays_per_subject <- icustays_tble %>%
  count(subject_id, name = "num_stays")

num_multiple_stays <- icu_stays_per_subject %>%
  filter(num_stays > 1) %>%
  nrow()

print(paste(num_multiple_stays))
```

```
[1] "16242"
```

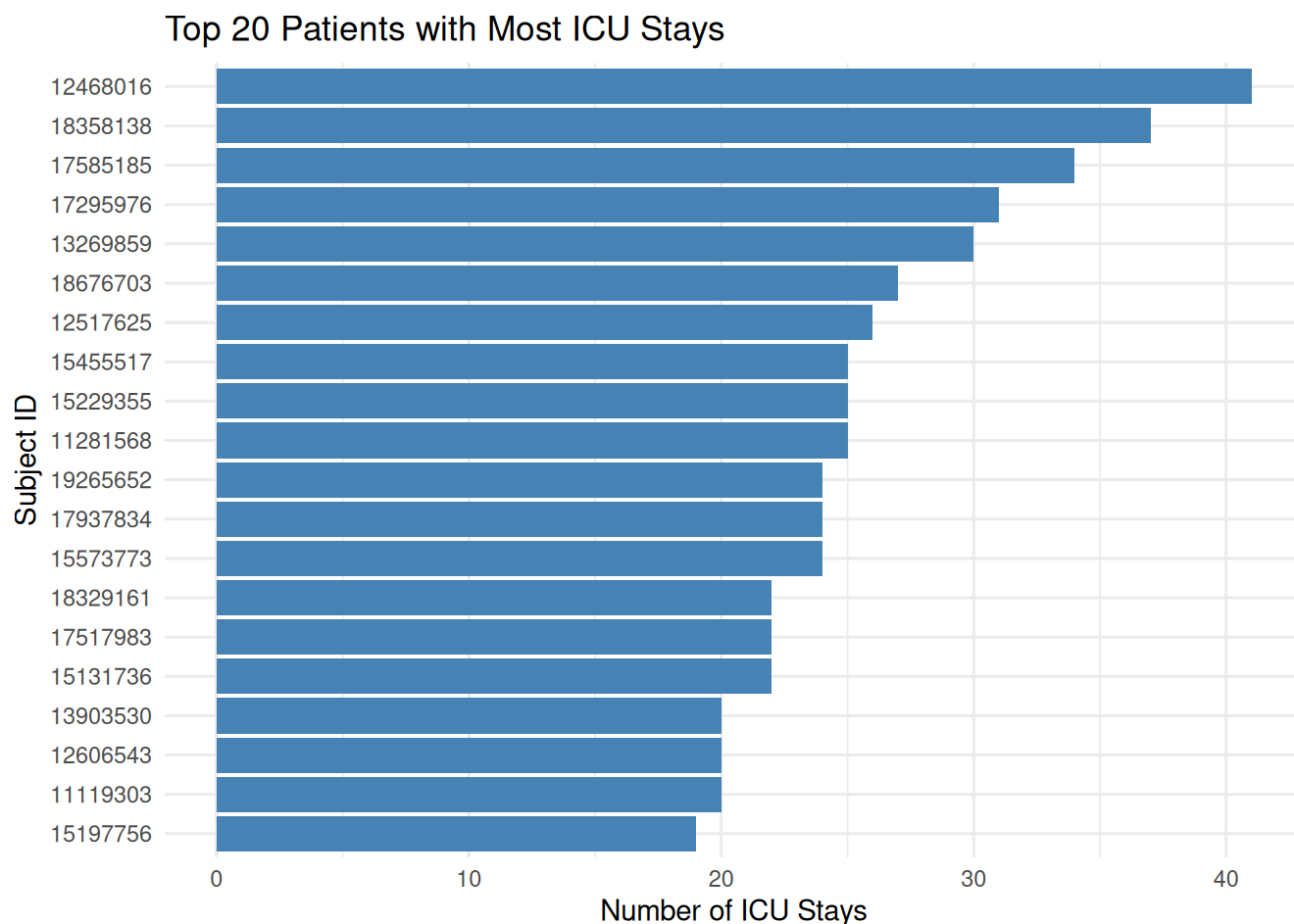
The number of ICU stays per 'subject_id' is shown through the following graphs:

1. Depicted below is a bar graph that displays the number of ICU stays for the top 20 patients with the most ICU stays. We can see that the maximum number of ICU stays amongst these 20 individuals ranges from between 19 to 41 ICU stays (including admission and discharge times). The general range of ICU stays is 1-41 stays.

```
icu_stays_summary <- icustays_tble %>%
  group_by(subject_id) %>%
  summarise(num_stays = n(), .groups = 'drop')

top_patients <- icu_stays_summary %>%
  arrange(desc(num_stays)) %>%
  head(20)

ggplot(top_patients, aes(x = reorder(subject_id, num_stays), y = num_stays)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(title = "Top 20 Patients with Most ICU Stays",
       x = "Subject ID",
       y = "Number of ICU Stays") +
  theme_minimal()
```



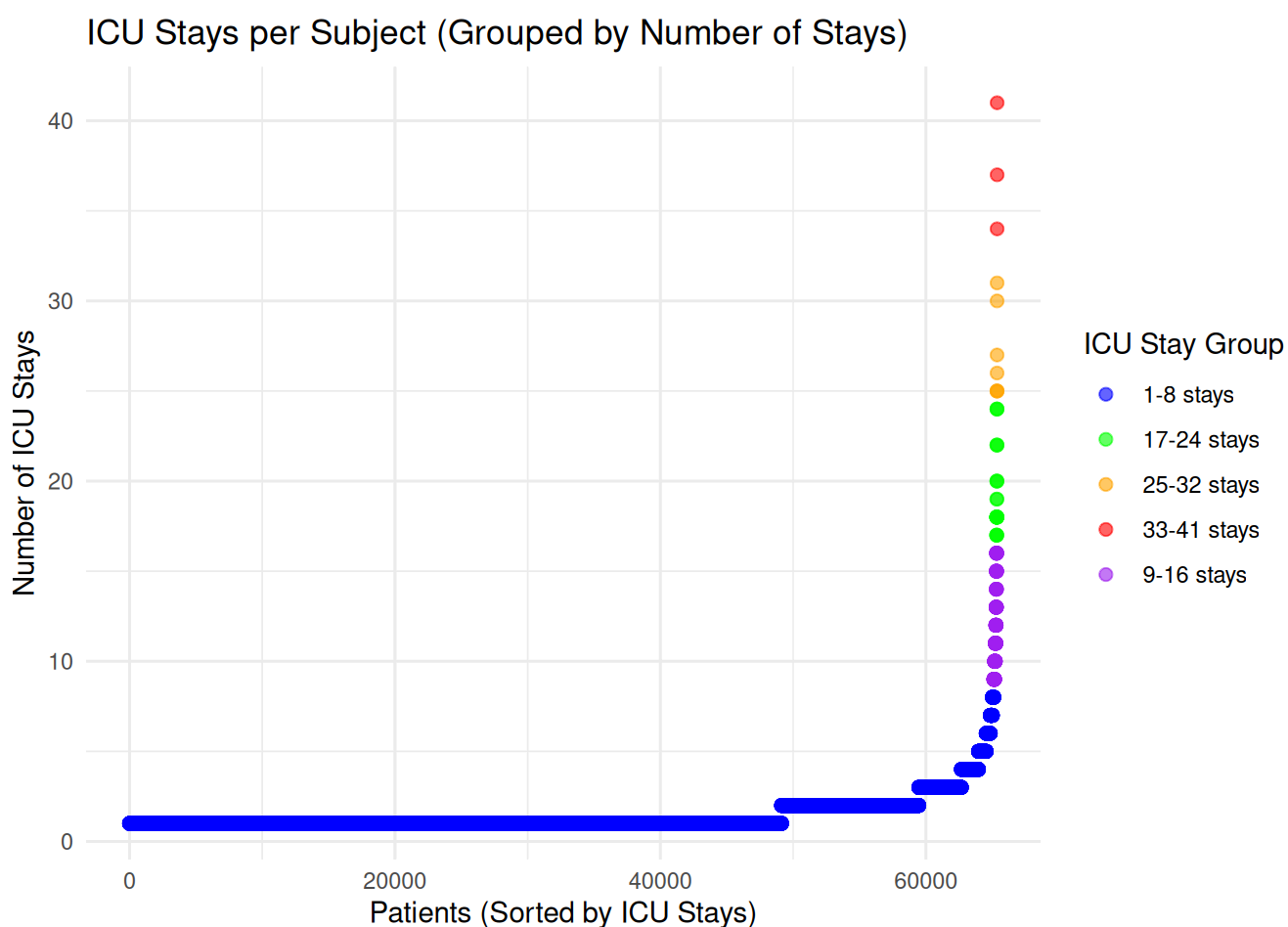
2. Below is a graph that visualizes the number of ICU stays per subject_id, with patients sorted in ascending order by the number of stays (each subject_id or patient is ranked starting from 1 depending on how many ICU stays they have). The visualization employs a color-coded scatter plot to distinguish different groups of ICU stays. Patients are grouped based on the number of ICU stays: 1. 1-8 stays (blue), 2. 9-16 stays (green), 3. 17-24 stays (orange), 4. 25-32 stays (red), 5. 33-41 stays (purple). From the scatter plot, we observe that the majority of points are blue, indicating that most patients fall within the 1-8 ICU stays range. This suggests that ICU stays are generally short for most patients. However, towards the right of the graph, the number of ICU stays increases steeply, forming a long-tail distribution. This pattern indicates that a small number of patients experience significantly higher ICU admissions. Overall, the distribution follows a right-skewed pattern, where most patients have a low ICU stay count, while a few patients have disproportionately high ICU stays.

```
icu_stays_summary <- icustays_tble %>%
  group_by(subject_id) %>%
  summarise(num_stays = n(), .groups = 'drop')

icu_stays_summary <- icu_stays_summary %>%
  mutate(stay_group = case_when(
    num_stays >= 1 & num_stays <= 8 ~ "1-8 stays",
    num_stays >= 9 & num_stays <= 16 ~ "9-16 stays",
    num_stays >= 17 & num_stays <= 24 ~ "17-24 stays",
    num_stays >= 25 & num_stays <= 32 ~ "25-32 stays",
    num_stays >= 33 & num_stays <= 41 ~ "33-41 stays"
  ))
```

```
icu_stays_summary <- icu_stays_summary %>%
  arrange(num_stays) %>%
  mutate(patient_index = row_number())

ggplot(icu_stays_summary, aes(x = patient_index,
                             y = num_stays,
                             color = stay_group)) +
  geom_point(alpha = 0.6, size = 2) +
  labs(title = "ICU Stays per Subject (Grouped by Number of Stays)",
       x = "Patients (Sorted by ICU Stays)",
       y = "Number of ICU Stays",
       color = "ICU Stay Group") +
  theme_minimal() +
  scale_color_manual(values = c("blue", "green", "orange", "red", "purple"))
```

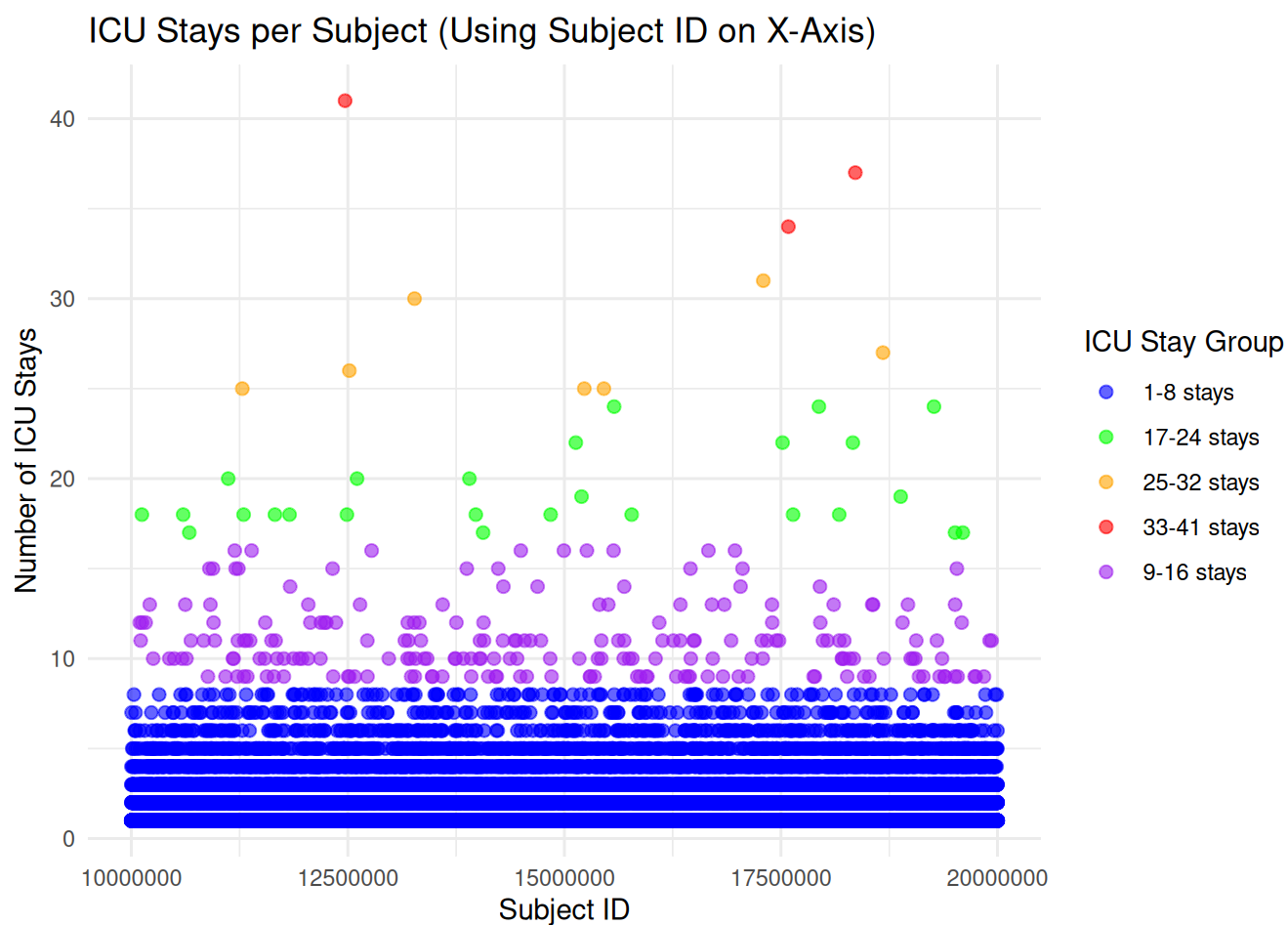


3. This graph provides similar information from the one above. The scatter plot visualizes the number of ICU stays per subject, using the specific id of each patient on the x-axis and the number of ICU stays on the y-axis. Each patient is categorized into different ICU stay groups, represented by distinct colors. The majority of patients have few ICU stays (1-8 stays) as depicted in blue. There is a small subset of patients that require frequent ICU visits as depicted in other colors other than blue.


```
icu_stays_summary <- icustays_tble %>%
  group_by(subject_id) %>%
  summarise(num_stays = n(), .groups = 'drop')

icu_stays_summary <- icu_stays_summary %>%
  mutate(stay_group = case_when(
    num_stays >= 1 & num_stays <= 8 ~ "1-8 stays",
    num_stays >= 9 & num_stays <= 16 ~ "9-16 stays",
    num_stays >= 17 & num_stays <= 24 ~ "17-24 stays",
    num_stays >= 25 & num_stays <= 32 ~ "25-32 stays",
    num_stays >= 33 & num_stays <= 41 ~ "33-41 stays"
  ))

ggplot(icu_stays_summary, aes(x = subject_id, y = num_stays, color = stay_group)) +
  geom_point(alpha = 0.6, size = 2) +
  labs(title = "ICU Stays per Subject (Using Subject ID on X-Axis)",
       x = "Subject ID",
       y = "Number of ICU Stays",
       color = "ICU Stay Group") +
  theme_minimal() +
  scale_color_manual(values = c("blue", "green", "orange", "red", "purple"))
```



Q3. admissions data

Information of the patients admitted into hospital is available in [admissions.csv.gz](https://mimic.mit.edu/docs/iv/modules/hosp/admissions/). 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_location,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.,P060TX,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,P39NW0,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`. **Done**

```
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

dtm (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.

```
print(admissions_tble)
```

```
# A tibble: 546,028 × 16
```

```
  subject_id  hadm_id admittime      disctime
      <dbl>    <dbl> <dtm>          <dtm>
1  10000032  22595853 2180-05-06 22:23:00 2180-05-07 17:15:00
2  10000032  22841357 2180-06-26 18:27:00 2180-06-27 18:49:00
3  10000032  25742920 2180-08-05 23:44:00 2180-08-07 17:50:00
4  10000032  29079034 2180-07-23 12:35:00 2180-07-25 17:55:00
5  10000068  25022803 2160-03-03 23:16:00 2160-03-04 06:26:00
6  10000084  23052089 2160-11-21 01:56:00 2160-11-25 14:52:00
7  10000084  29888819 2160-12-28 05:11:00 2160-12-28 16:07:00
8  10000108  27250926 2163-09-27 23:17:00 2163-09-28 09:04:00
9  10000117  22927623 2181-11-15 02:05:00 2181-11-15 14:52:00
10 10000117  27988844 2183-09-18 18:10:00 2183-09-21 16:30:00
```

```
# i 546,018 more rows
```

```
# i 12 more variables: deathtime <dtm>, admission_type <chr>,
# admit_provider_id <chr>, admission_location <chr>,
# discharge_location <chr>, insurance <chr>, language <chr>,
# marital_status <chr>, race <chr>, edregtime <dtm>, edouttime <dtm>,
# hospital_expire_flag <dbl>
```

(3.1 Solution) The admission_tble tibble has 546,028 rows and 16 columns.

Q3.2 Summary and visualization

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.

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

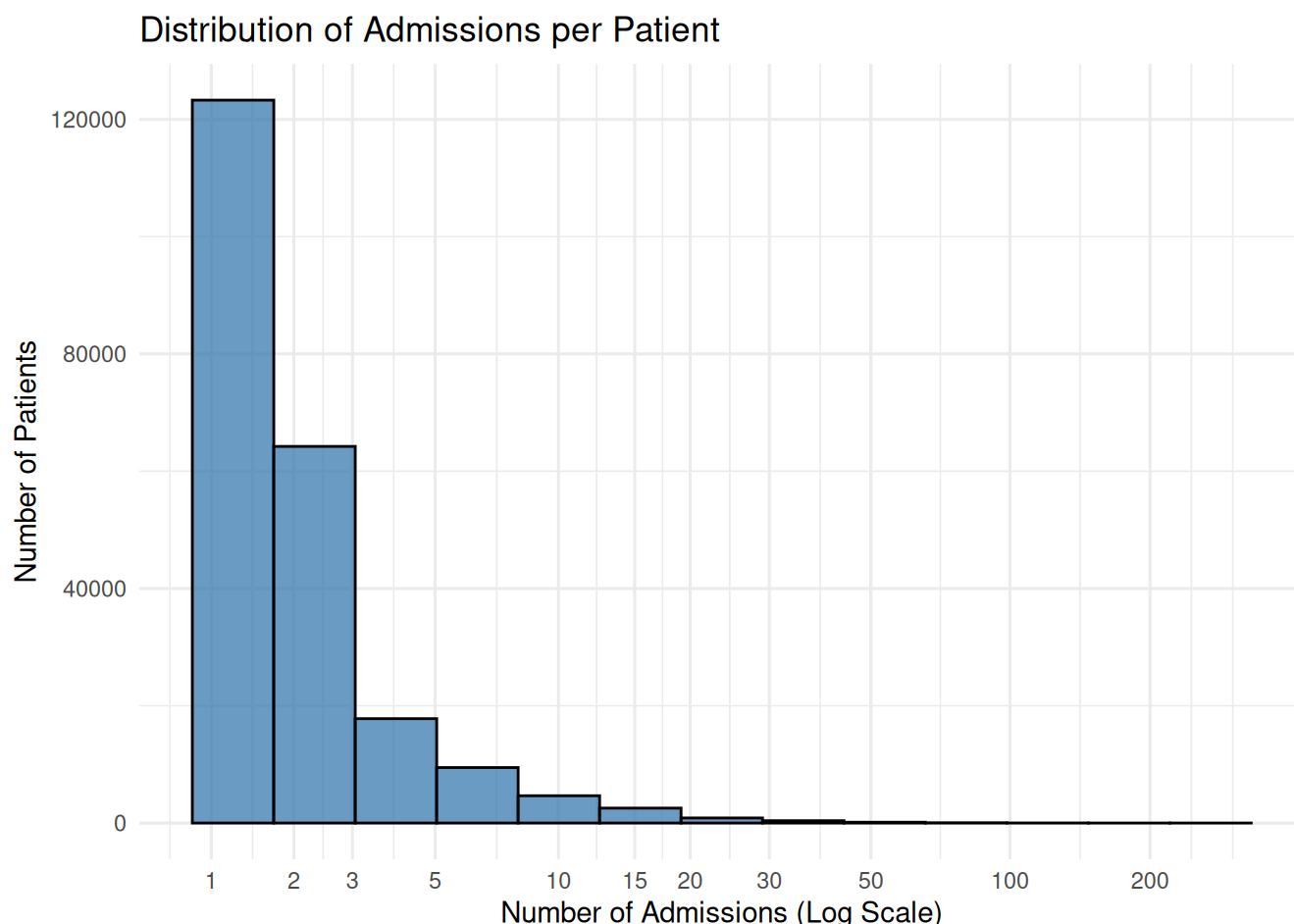
1. Number of admissions per patient

Analysis: The histogram illustrates the distribution of hospital admissions per patient on a log scale. The x-axis (log-transformed) represents the number of admissions per patient, while the y-axis represents the number of patients. From the histogram, we observe that the majority of patients have only 1-3 admissions, indicating that most individuals require minimal hospital visits. However, a small subset of patients exhibit significantly higher admission counts, with some exceeding 10+ admissions. Notably, a few extreme cases surpass 200 admissions, highlighting a group of patients with frequent and recurrent hospitalizations.

```
admissions_summary <- admissions_tble %>%
  group_by(subject_id) %>%
  summarise(num_admissions = n(), .groups = 'drop')

ggplot(admissions_summary, aes(x = num_admissions)) +
```

```
geom_histogram(binwidth = 0.4, fill = "steelblue",
               color = "black", alpha = 0.8) +
scale_x_continuous(trans = "log1p",
                   breaks = c(1, 2, 3, 5, 10, 15, 20, 30, 50, 100, 200)) +
labs(title = "Distribution of Admissions per Patient",
     x = "Number of Admissions (Log Scale)",
     y = "Number of Patients") +
theme_minimal()
```



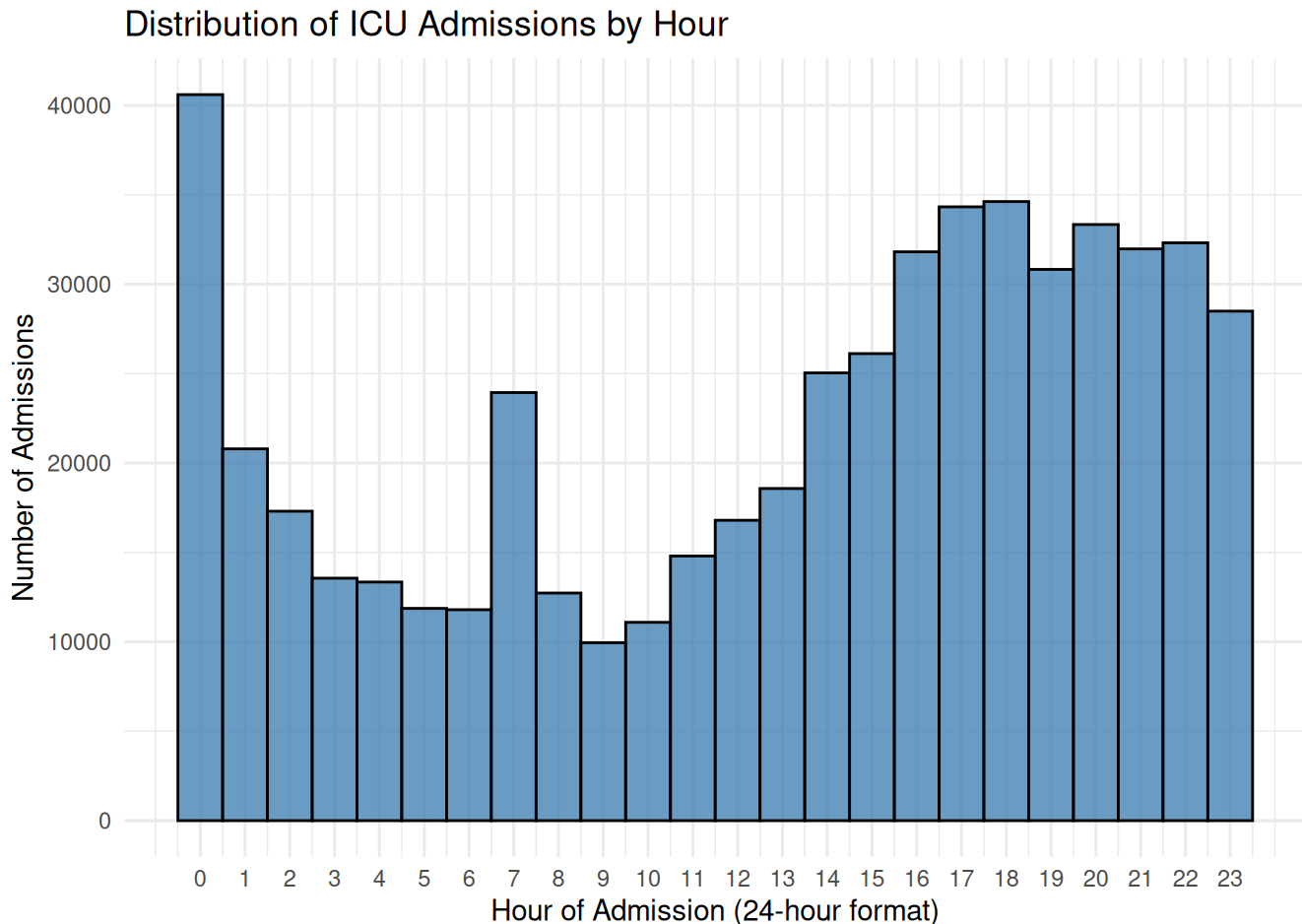
2. Admission hour (anything unusual?)

Analysis There is a high sharp spike at Midnight (00:00) which is likely not a true clinical pattern but rather an artifact of hospital data entry practices, such as batch processing of admissions at the start of a new day. The second peak occurs around 7-8 AM which may reflect morning clinical assessments leading to ICU transfers. Admissions then gradually rise throughout the day, peaking between 4-8 PM.

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

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

```
y = "Number of Admissions") +  
theme_minimal()
```



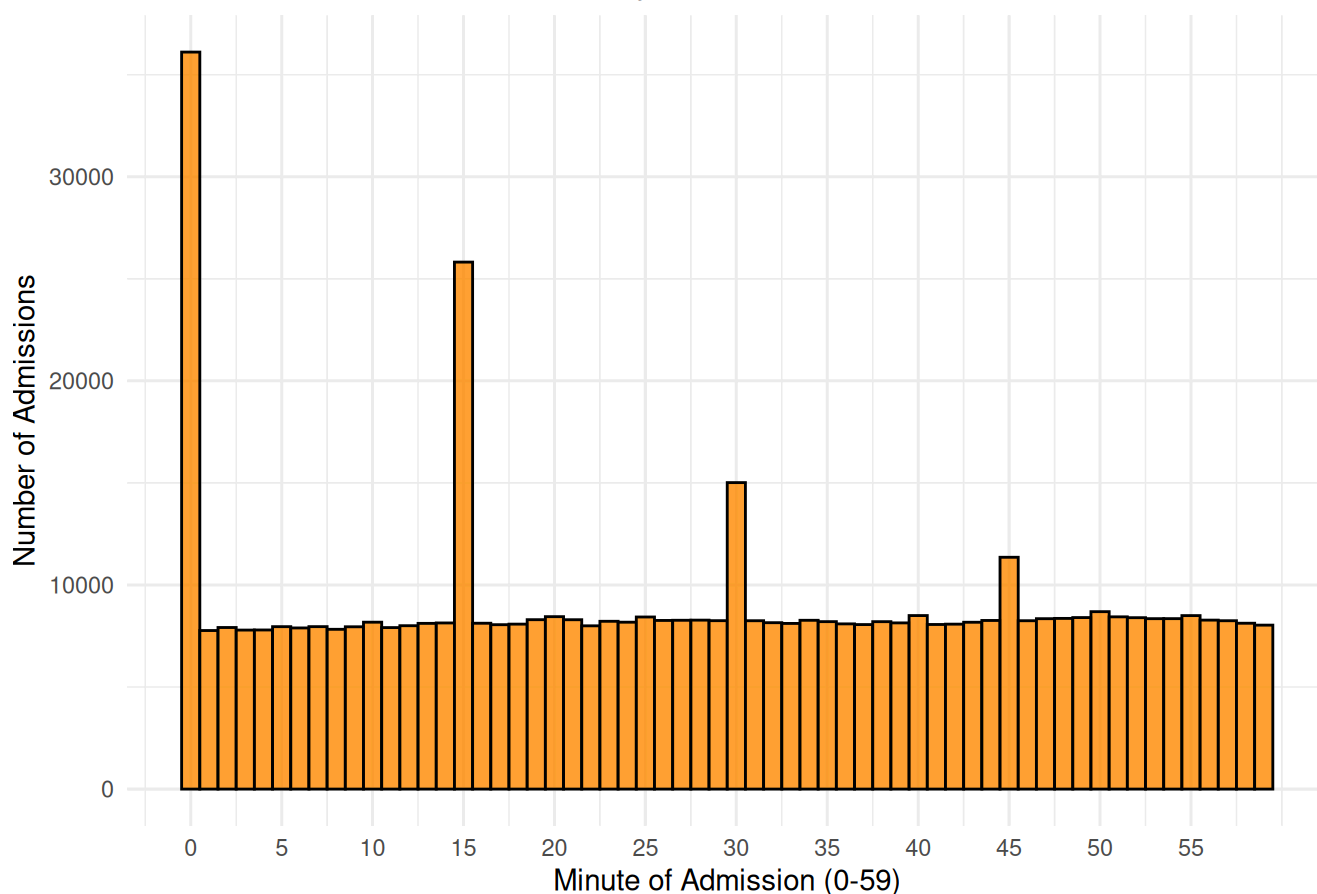
3. Admission minute (anything unusual?)

Analysis: The histogram displays the distribution of ICU admission by the minute of the hour (0-59). There are sharp pikes in admissions at minutes 0, 15, 30, and 45 minutes, which are significantly higher than the other minutes. This suggests that admissions are often recorded at rounded times rather than exact times when patients arrive. This pattern is unlikely because there would be natural variations in patient arrivals. Instead, it could be a result of the hospital staff rounding admission times when entering data. Apart from the spikes, the distribution across other minutes is relatively uniform, signifying that the real admission times likely occur throughout the hour but are being recorded in a biased manner.

```
admissions_tble <- admissions_tble %>%  
  mutate(admit_minute = minute(admittime))  
  
ggplot(admissions_tble, aes(x = admit_minute)) +  
  geom_histogram(binwidth = 1, fill = "darkorange",  
                 color = "black", alpha = 0.8) +  
  scale_x_continuous(breaks = seq(0, 59, by = 5)) +  
  labs(title = "Distribution of ICU Admissions by Minute",  
       x = "Minute of Admission (0-59)",
```

```
y = "Number of Admissions") +
theme_minimal()
```

Distribution of ICU Admissions by Minute



4. Length of hospital stay (from admission to discharge) (anything unusual?)

Analysis: The histogram that evaluates the hospital length of stay distribution showcases a heavily right skewed distribution where most patients have very short hospital stays, with the highest frequency at 1 day. The majority of admissions are under 5 days and there is a sharp decline in the number of admissions as the length of stay increases. We are able to see that some patients have long hospitalizations (up to 50 days), but these cases are much fewer. The long tail suggests that a small subset of patients require extended hospital care.

```
admissions_tble <- admissions_tble %>%
  mutate(
    admit_time = as.POSIXct(admittime,
                           format = "%Y-%m-%d %H:%M:%S", tz = "UTC"),
    discharge_time = as.POSIXct(dischtime,
                               format = "%Y-%m-%d %H:%M:%S", tz = "UTC")
  ) %>%
  mutate(LOS_days = as.numeric(
    difftime(discharge_time, admit_time, units = "days")
  ))

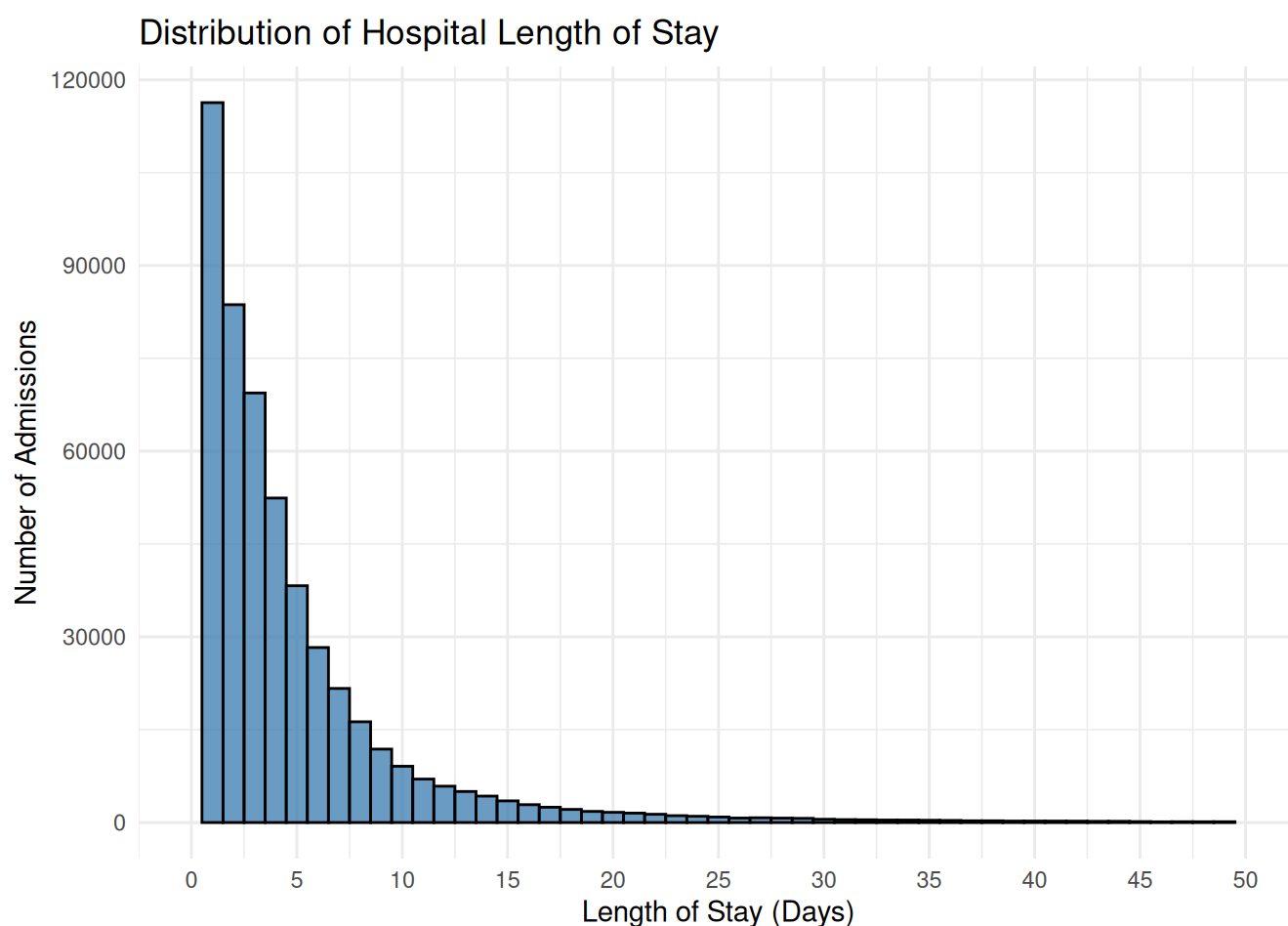
admissions_tble <- admissions_tble %>%
```

```
filter(LOS_days >= 0, LOS_days < 365)

ggplot(admissions_tble, aes(x = LOS_days)) +
  geom_histogram(binwidth = 1, fill = "steelblue",
                 color = "black", alpha = 0.8) +
  scale_x_continuous(
    breaks = seq(0, 50, by = 5),
    limits = c(0, 50) # Limit x-axis to 50 days
  ) +
  labs(title = "Distribution of Hospital Length of Stay",
       x = "Length of Stay (Days)",
       y = "Number of Admissions") +
  theme_minimal()
```

Warning: Removed 1929 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](https://mimic.mit.edu/docs/iv/modules/hosp/patients/). 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/) (<https://mimic.mit.edu/docs/iv/modules/hosp/patients/>) as a tibble `patients_tble`.

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

Rows: 364627 Columns: 6

— Column specification —

Delimiter: ","

chr (2): gender, anchor_year_group

dbl (3): subject_id, anchor_age, anchor_year

date (1): dod

i Use ``spec()`` to retrieve the full column specification for this data.

i Specify the column types or set ``show_col_types = FALSE`` to quiet this message.

```
print(patients_tble)
```

A tibble: 364,627 × 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
7	10000108	M	25	2163	2014 - 2016	NA
8	10000115	M	24	2154	2017 - 2019	NA
9	10000117	F	48	2174	2008 - 2010	NA
10	10000161	M	60	2163	2020 - 2022	NA

i 364,617 more rows

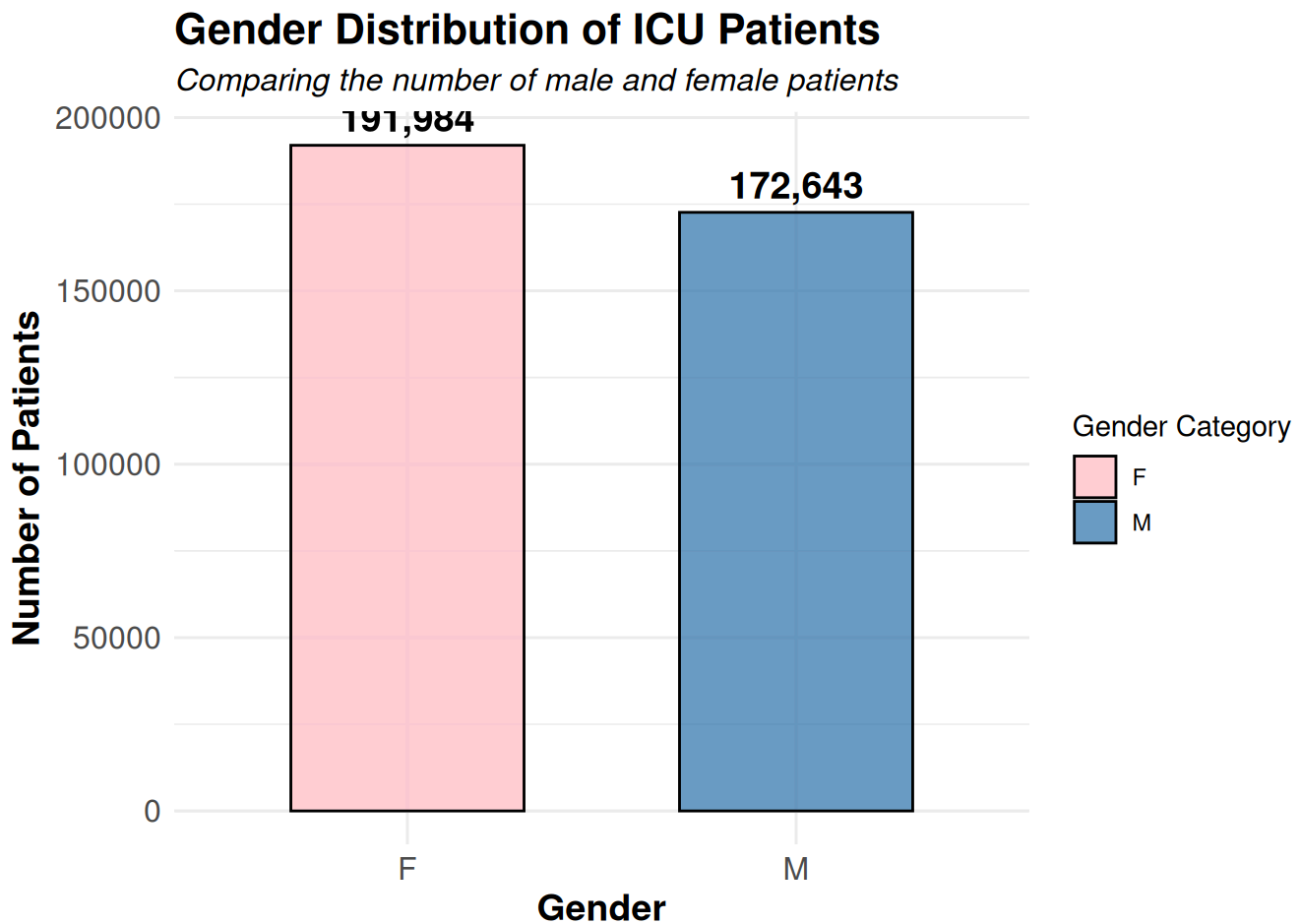
Q4.2 Summary and visualization

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

Summary Graph of Gender (genotypical sex of the patient): The bar chart illustrates the gender distribution of patients, showing a slight female majority. While females are more frequent than males, the difference is not substantial, indicating a fairly balanced representation of both genders in the dataset. More specifically, the number of females in the `patients_tble` is 191,984 and the number of males in the `patients_tble` is 172,643.

```
ggplot(patients_tble, aes(x = gender, fill = gender)) +
  geom_bar(color = 'black', alpha = 0.8, width = 0.6) +
  scale_fill_manual(values = c("M" = "steelblue", "F" = "pink")) +
  labs(
    title = "Gender Distribution of ICU Patients",
    subtitle = "Comparing the number of male and female patients",
    x = "Gender",
    y = "Number of Patients",
    fill = "Gender Category"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 16, face = "bold"),
    plot.subtitle = element_text(size = 12, face = "italic"),
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12),
    axis.title = element_text(size = 14, face = "bold"),
    legend.position = "right"
  ) +
  geom_text(stat = "count", aes(label = scales::comma(..count..)),
    vjust = -0.5, size = 5, fontface = "bold")
```

Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.
 i Please use ``after_stat(count)`` instead.

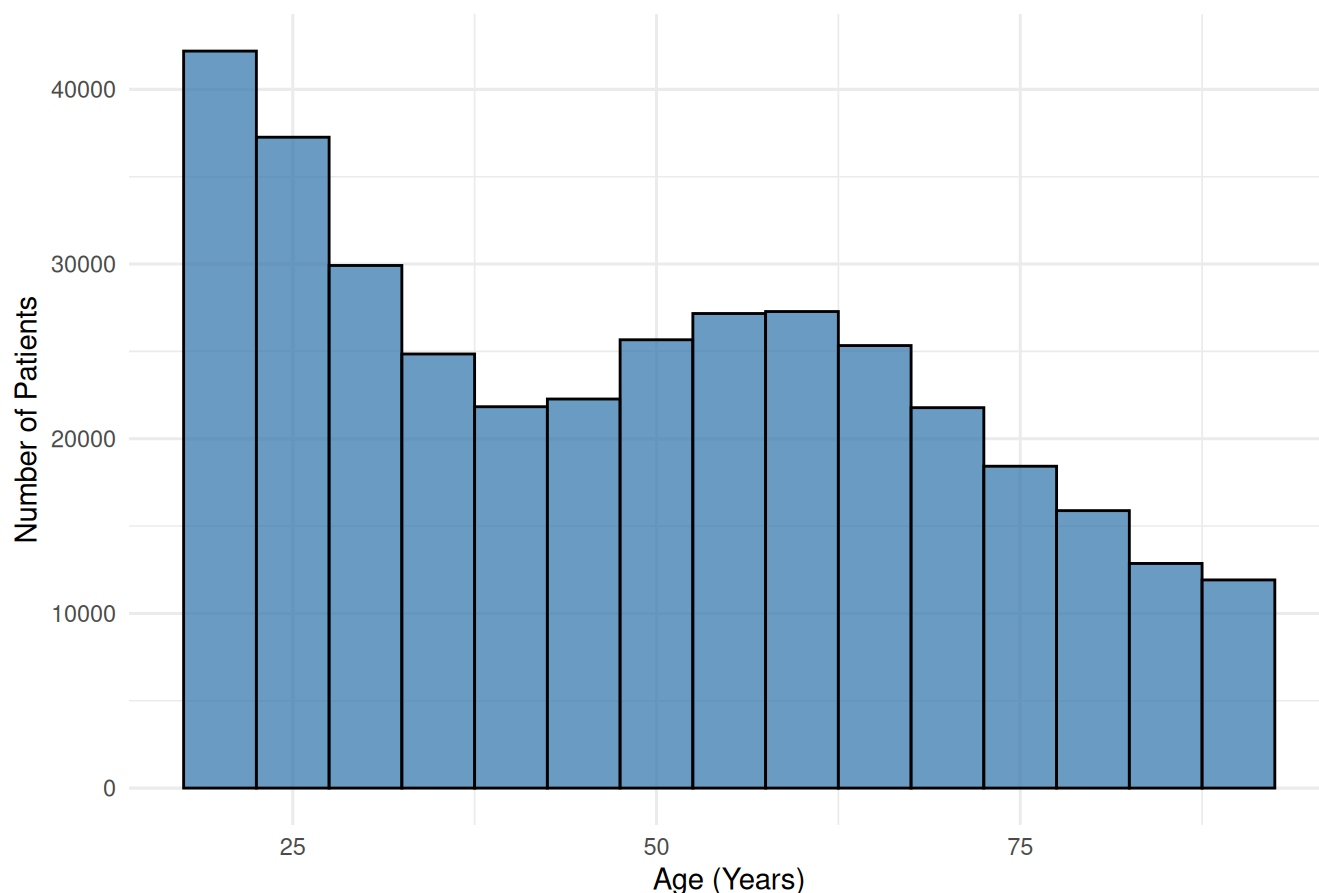


Summary Graph of Anchor_Age:

Anchor_age represents a patient's age in the anchor_year. If a patient is over 89 years old, their anchor_age is set to 91, regardless of their actual age. The histogram depicts the age distribution of patients, showing a right-skewed pattern. Younger patients (under 25) make up the largest group. The number of patients gradually declines with age, though there is a slight increase around ages 45-65. Fewer patients are observed in older age groups (75+ years), which aligns with natural aging and mortality trends.

```
# 2. Histogram: Age Distribution
ggplot(patients_tble, aes(x = anchor_age)) +
  geom_histogram(binwidth = 5, fill = "steelblue",
                 color = "black", alpha = 0.8) +
  labs(title = "Age Distribution of Patients",
       x = "Age (Years)",
       y = "Number of Patients") +
  theme_minimal()
```

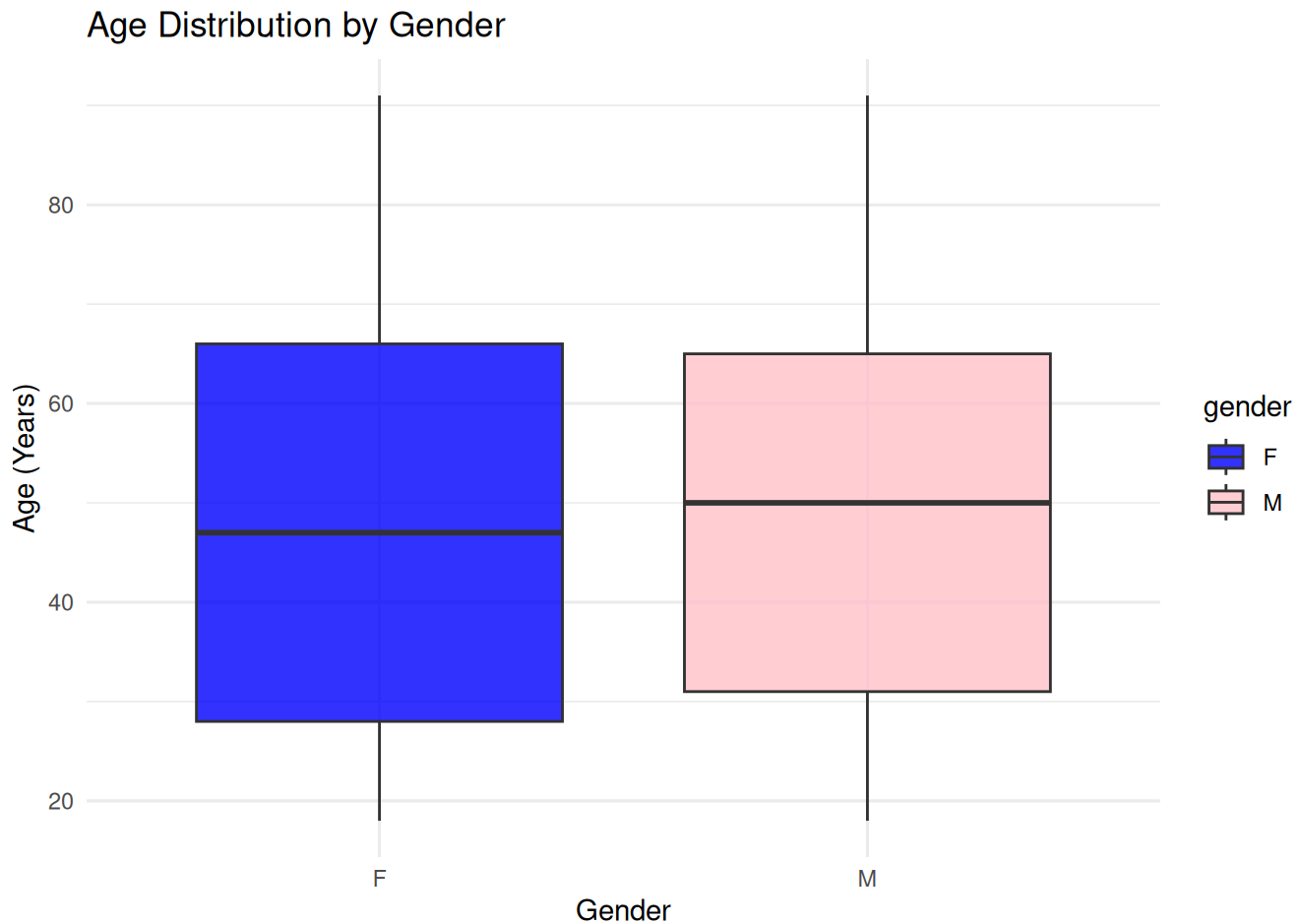
Age Distribution of Patients



Visualizing the 2 variables together: Gender and Anchor_Age:

The box plot visualizes the age distribution of patients by genders, females and males. The median age for both genders appears to be similar, around the mid-50s. This suggests that the age demographics of male and female patients are comparable. The IQR (middle 50 percent of values) spans from approximately 30 to 70 years for both genders. This indicates that most ICU patients fail within this age range. There is no significant visual difference in the age distribution between males and females which suggests that both genders have a similar age profile in ICU admissions.

```
# 3. Box Plot: Age by Gender
ggplot(patients_tble, aes(x = gender, y = anchor_age, fill = gender)) +
  geom_boxplot(alpha = 0.8) +
  labs(title = "Age Distribution by Gender",
       x = "Gender",
       y = "Age (Years)") +
  theme_minimal() +
  scale_fill_manual(values = c("blue", "pink"))
```



Q5. Lab results

[labevents.csv.gz](https://mimic.mit.edu/docs/iv/modules/hosp/labevents/) (<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,value_num,value_uom,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.

```
library(duckdb)
parquet_file <- "labevents_pq/part-0.parquet"
start_time <- Sys.time()
arrow_data <- open_dataset(parquet_file, format = "parquet")
duckdb_conn <- dbConnect(duckdb::duckdb())

invisible(to_duckdb(
  .data = arrow_data,
  con = duckdb_conn,
  table_name = "labevents",
  auto_disconnect = FALSE
))
```

```
subset_labevents <- tbl(duckdb_conn, "labevents") %>%
  filter(itemid %in% c(50912, 50971, 50983, 50902,
                     50882, 51221, 51301, 50931)) %>%
  arrange(subject_id, charttime, itemid,) %>%
  collect()
dbDisconnect(duckdb_conn)
```

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

labevents_filtered <- subset_labevents %>%
  inner_join(icustays_tble, by = "subject_id") %>%
  filter(storetime < intime)
```

Warning in inner_join(., icustays_tble, by = "subject_id"): Detected an unexpected many-to-many relationship between `x` and `y`.

- i Row 2341 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.

```
labevents_filtered <- labevents_filtered %>%
  mutate(valuenum = as.numeric(valuenum))

labevents_last <- labevents_filtered %>%
  group_by(subject_id, stay_id, itemid) %>%
  slice_max(order_by = storetime, n = 1, with_ties = FALSE) %>%
  ungroup()

labevents_tble <- labevents_last %>%
  select(subject_id, stay_id, itemid, valuenum) %>%
  pivot_wider(names_from = itemid, values_from = valuenum)

colnames(labevents_tble) <- c(
  "subject_id", "stay_id", "bicarbonate", "chloride", "creatinine",
  "glucose", "potassium", "sodium", "hematocrit", "wbc"
)

labevents_tble <- labevents_tble %>%
  mutate(across(where(is.list), ~ as.numeric(unlist(.))))

cat("Data processing complete.\n")
```

Data processing complete.

```
print(head(labevents_tble))
```

```
# A tibble: 6 × 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>

Q6. Vitals from charted events

[chartevents.csv.gz](https://mimic.mit.edu/docs/iv/modules/icu/chartevents/) (<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](https://mimic.mit.edu/docs/iv/modules/icu/chartevents/) are

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

```
subject_id,hadm_id,stay_id,caregiver_id,charttime,storetime,itemid,value,valuenum,valueuom,warning
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/) (https://mimic.mit.edu/docs/iv/modules/icu/d_items/) is the dictionary for the `itemid` in [chartevents.csv.gz](https://mimic.mit.edu/docs/iv/modules/icu/chartevents/).

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

```
itemid,label,abbreviation,linksto,category,unitname,param_type,lownormalvalue,highnormalv
alue
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.



Solution Note that the subset_chartevents contains information regarding diastolic non-invasive blood pressure 220180 and not 220181 (from last homework).

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

chartevents_filtered <- subset_chartevents %>%
  inner_join(icustays_tble, by = "stay_id") %>%
  filter(storetime >= intime & storetime < outtime)

chartevents_filtered <- chartevents_filtered %>%
  rename(subject_id = subject_id.x) %>%
  select(-subject_id.y)

first_store_per_stay <- chartevents_filtered %>%
  group_by(subject_id, stay_id, itemid) %>%
  arrange(storetime) %>%
  slice_min(order_by = storetime, n = 1) %>%
  ungroup()

chartevents_filtered <- chartevents_filtered %>%
  inner_join(first_store_per_stay %>%
    select(subject_id, stay_id, itemid, storetime),
    by = c("subject_id", "stay_id", "itemid", "storetime")
  ) %>%
  select(-storetime)
```

Warning in inner_join(., first_store_per_stay %>% select(subject_id, stay_id, : Detected
 an unexpected many-to-many relationship between `x` and `y`.
 i Row 44 of `x` matches multiple rows in `y`.

- i Row 6 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.

```

chartevents_filtered <- chartevents_filtered %>%
  group_by(subject_id, stay_id, itemid) %>%
  summarize(valuenum_avg = mean(valuenum, na.rm = TRUE),
    .groups = "drop") %>%
  ungroup()

chartevents_tble <- chartevents_filtered %>%
  pivot_wider(
    names_from = itemid, values_from = valuenum_avg,
    names_prefix = "vital_"
  )

chartevents_tble <- chartevents_tble %>%
  rename(
    heart_rate = vital_220045,
    non_invasive_blood_pressure_systolic = vital_220179,
    non_invasive_blood_pressure_diastolic = vital_220180,
    temperature_fahrenheit = vital_223761,
    respiratory_rate = vital_220210
  ) %>%
  arrange(subject_id, stay_id)

print(chartevents_tble)

```

A tibble: 94,363 × 7

	subject_id	stay_id	heart_rate	non_invasive_blood_pr... ¹	non_invasive_blood_p... ²
	<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	10000032	39553978	91	84	48
2	10000690	37081114	78	106	56.5
3	10000980	39765666	76	154	102
4	10001217	34592300	79.3	156	93.3
5	10001217	37067082	86	151	90
6	10001725	31205490	86	73	56
7	10001843	39698942	124.	110	78
8	10001884	37510196	49	174.	30.5
9	10002013	39060235	80	98.5	62
10	10002114	34672098	110.	112	80

i 94,353 more rows

i abbreviated names: ¹non_invasive_blood_pressure_systolic,

²non_invasive_blood_pressure_diastolic

i 2 more variables: respiratory_rate <dbl>, temperature_fahrenheit <dbl>

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.

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.



```
icustays_age <- icustays_tble %>%
  mutate(intime_year = year(as.Date(intime)))

age_at_intime <- icustays_age %>%
  left_join(
    patients_tble %>%
      select(subject_id, anchor_age, anchor_year),
    by = "subject_id"
  ) %>%
  mutate(age_at_intime = anchor_age + (intime_year - anchor_year)) %>%
  select(subject_id, stay_id, age_at_intime)

icustays_tble <- icustays_tble %>%
  left_join(age_at_intime, by = c("subject_id", "stay_id"))
```

```
icustays_filtered <- icustays_tble %>%
  inner_join(
    patients_tble %>%
      select(
        subject_id, anchor_age, anchor_year,
        anchor_year_group, dod, gender
      ),
    by = "subject_id"
  ) %>%
  filter(anchor_age >= 18)

icu_admissions <- icustays_filtered %>%
  left_join(
    admissions_tble %>%
      select(-admit_hour, -admit_minute),
    by = c("subject_id", "hadm_id")
  )

icu_vitals <- icu_admissions %>%
```

```

left_join(chartevents_tble, by = c("subject_id", "stay_id"))

icu_final <- icu_vitals %>%
  left_join(labevents_tble, by = c("subject_id", "stay_id"))

mimic_icu_cohort <- icu_final %>%
  select(-LOS_days, -admit_time, -discharge_time) %>% # Exclude unwanted columns
  arrange(subject_id, hadm_id, stay_id) %>%
  distinct()

print(mimic_icu_cohort)

```

A tibble: 94,458 × 41

```

  subject_id  hadm_id  stay_id first_careunit last_careunit intime
    <dbl>    <dbl>    <dbl> <chr>          <chr>          <dtm>
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 35 more variables: outtime <dtm>, los <dbl>, age_at_intime <dbl>,
#   anchor_age <dbl>, anchor_year <dbl>, anchor_year_group <chr>, dod <date>,
#   gender <chr>, admittime <dtm>, disctime <dtm>, deathtime <dtm>,
#   admission_type <chr>, admit_provider_id <chr>, admission_location <chr>,
#   discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dtm>, edouttime <dtm>, ...

```

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)

The summary statistics for the individual variables `los`, `race`, `insurance`, `marital_status`, `gender`, and `age at intime` are stated below:

```

# Summary statistics for individual variables
summary_stats <- mimic_icu_cohort %>%
  summarise(
    los_mean = mean(los, na.rm = TRUE),
    los_median = median(los, na.rm = TRUE),
    los_sd = sd(los, na.rm = TRUE),

```

```

    los_min = min(los, na.rm = TRUE),
    los_max = max(los, na.rm = TRUE)
  )

# Summary of categorical variables
categorical_summary <- mimic_icu_cohort %>%
  summarise(
    race_levels = n_distinct(race),
    insurance_levels = n_distinct(insurance),
    marital_status_levels = n_distinct(marital_status),
    gender_levels = n_distinct(gender)
  )

# Summary of numerical variable (Age at ICU admission)
age_intime_summary <- mimic_icu_cohort %>%
  summarise(
    age_mean = mean(age_at_intime, na.rm = TRUE),
    age_median = median(age_at_intime, na.rm = TRUE),
    age_sd = sd(age_at_intime, na.rm = TRUE),
    age_min = min(age_at_intime, na.rm = TRUE),
    age_max = max(age_at_intime, na.rm = TRUE)
  )

# Display all summaries in a readable format
cat("\n==== Length of ICU Stay (LOS) Summary =====\n")

```

==== Length of ICU Stay (LOS) Summary =====

```
cat(paste0("Mean: ", summary_stats$los_mean, "\n"))
```

Mean: 3.63002485183264

```
cat(paste0("Median: ", summary_stats$los_median, "\n"))
```

Median: 1.96564814814815

```
cat(paste0("Std Dev: ", summary_stats$los_sd, "\n"))
```

Std Dev: 5.40247353755194

```
cat(paste0("Min: ", summary_stats$los_min, "\n"))
```

Min: 0.00125

```
cat(paste0("Max: ", summary_stats$los_max, "\n"))
```

Max: 226.403078703704

```
cat("\n==== Categorical Variables Summary =====\n")
```

==== Categorical Variables Summary =====

```
cat(paste0("Race Categories: ", categorical_summary$race_levels, "\n"))
```

Race Categories: 34

```
cat(paste0("Insurance Types: ", categorical_summary$insurance_levels, "\n"))
```

Insurance Types: 6

```
cat(paste0("Marital Status Levels: ",  
          categorical_summary$marital_status_levels, "\n"))
```

Marital Status Levels: 5

```
cat(paste0("Gender Levels: ", categorical_summary$gender_levels, "\n"))
```

Gender Levels: 2

```
cat("\n==== Age at Intime Summary =====\n")
```

==== Age at Intime Summary =====

```
cat(paste0("Mean: ", age_intime_summary$age_mean, "\n"))
```

Mean: 64.7862224480722

```
cat(paste0("Median: ", age_intime_summary$age_median, "\n"))
```

Median: 66

```
cat(paste0("Std Dev: ", age_intime_summary$age_sd, "\n"))
```

Std Dev: 16.7395268673104

```
cat(paste0("Min: ", age_intime_summary$age_min, "\n"))
```

Min: 18

```
cat(paste0("Max: ", age_intime_summary$age_max, "\n"))
```

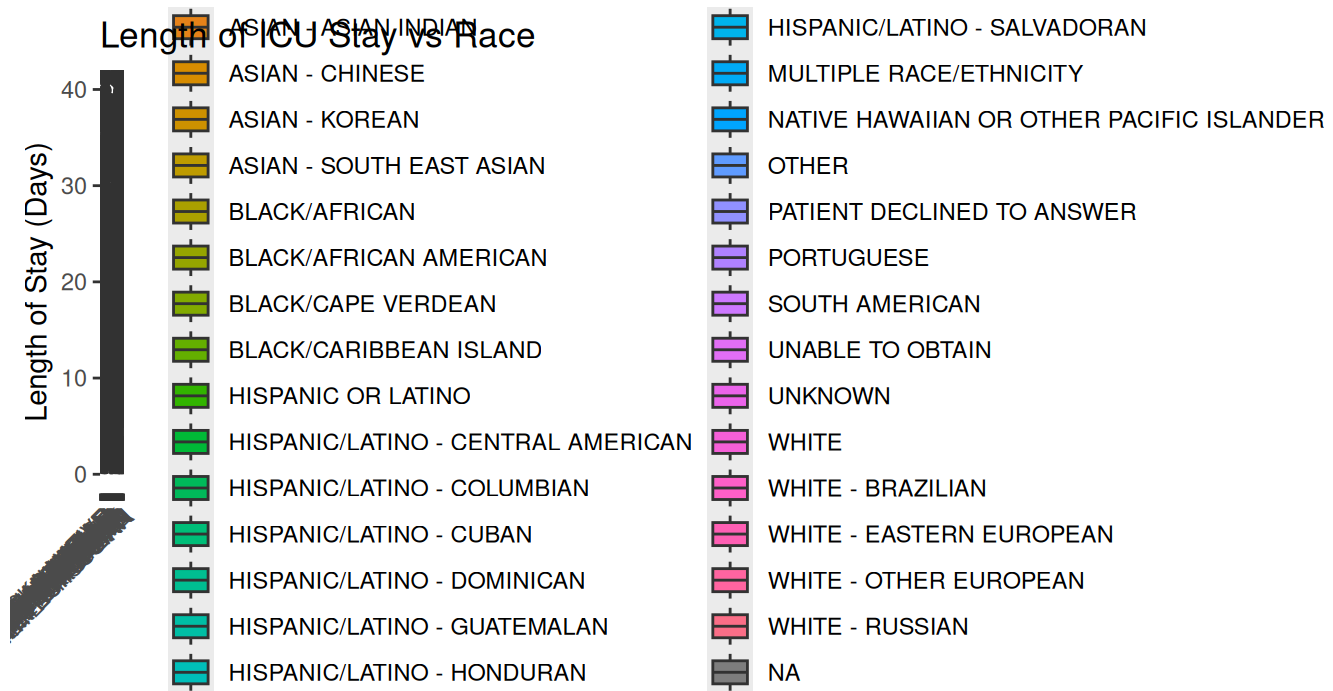
Max: 103

1. Boxplot of LOS (Length of ICU Stay) versus Race:

We evaluate the relationship length of stay for the patients for the given ICU stay and race through multiple histograms. We are able to see that most racial groups have a median ICU stay of a few days, indicating that ICU stays are typically short for the majority of patients. There are many outliers across all racial groups, with some patients staying over 40 days. Certain groups, such as “White - Eastern European” and “Other” appear to have wider range of ICU stays. Groups like “Hispanic/Latino - Central American” and “Black/African American” show slightly higher variability compared to others. The N/A category (potentially missing race data) has a significant spread, which may indicate data inconsistencies. The distribution suggests most ICU stays are short, but a small subset of patients experience prolonged ICU admissions. Note that we limited the y-axis to 40 days (length of stay) for better readability of the graph.

```
# Boxplot: LOS vs Race
ggplot(mimic_icu_cohort, aes(x = race, y = los, fill = race)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Length of ICU Stay vs Race",
       x = "Race",
       y = "Length of Stay (Days)") +
  coord_cartesian(ylim = c(0, 40))
```

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



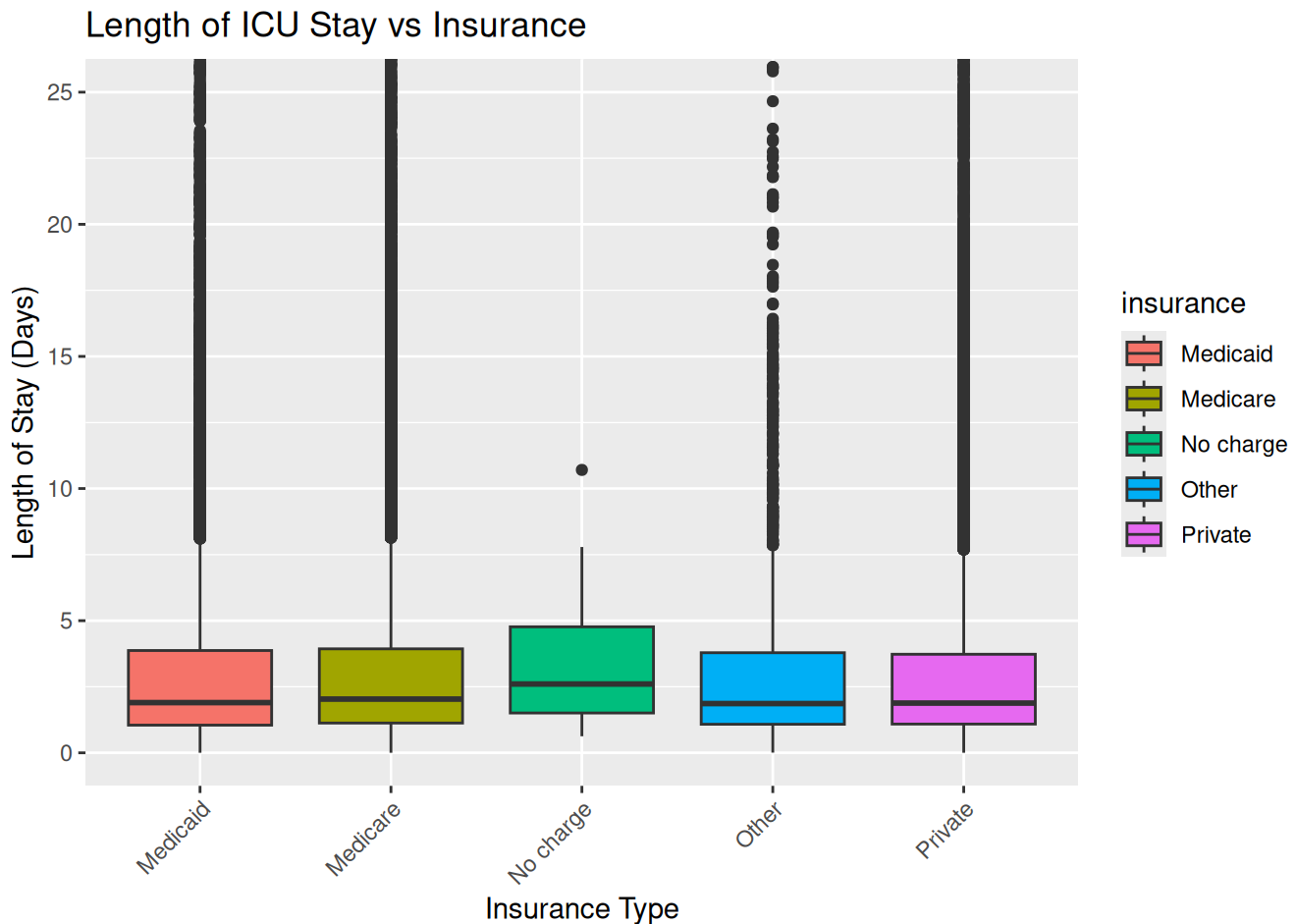
Race

2. Boxplot of LOS versus Insurance

The boxplots below visualize the distribution of ICU length of stay (LOS) across different insurance types. Most insurance groups have a median ICU stay of a few days, with Medicare and Medicaid patients showing slightly longer median stays compared to other groups. All insurance categories exhibit significant variability in LOS. The “Other” and “Private” insurance groups display a wider spread, indicating a more diverse range of ICU stay durations. Outliers extend beyond 25 days in every insurance category, suggesting that while most patients have relatively short ICU stays, a small subset experiences extended admissions. “No charge” patients show slightly more variability than Medicaid and Medicare, possibly indicating that patients under financial assistance programs experience more variable ICU stays, potentially due to socioeconomic or healthcare access factors. Note: We limit the y-axis (length of stay in days) to 25 days for better readability of the graphs.

```
ggplot(mimic_icu_cohort %>% filter(!is.na(insurance)),
       aes(x = insurance, y = los, fill = insurance)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Length of ICU Stay vs Insurance",
       x = "Insurance Type",
       y = "Length of Stay (Days)") +
  coord_cartesian(ylim = c(0, 25))
```

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

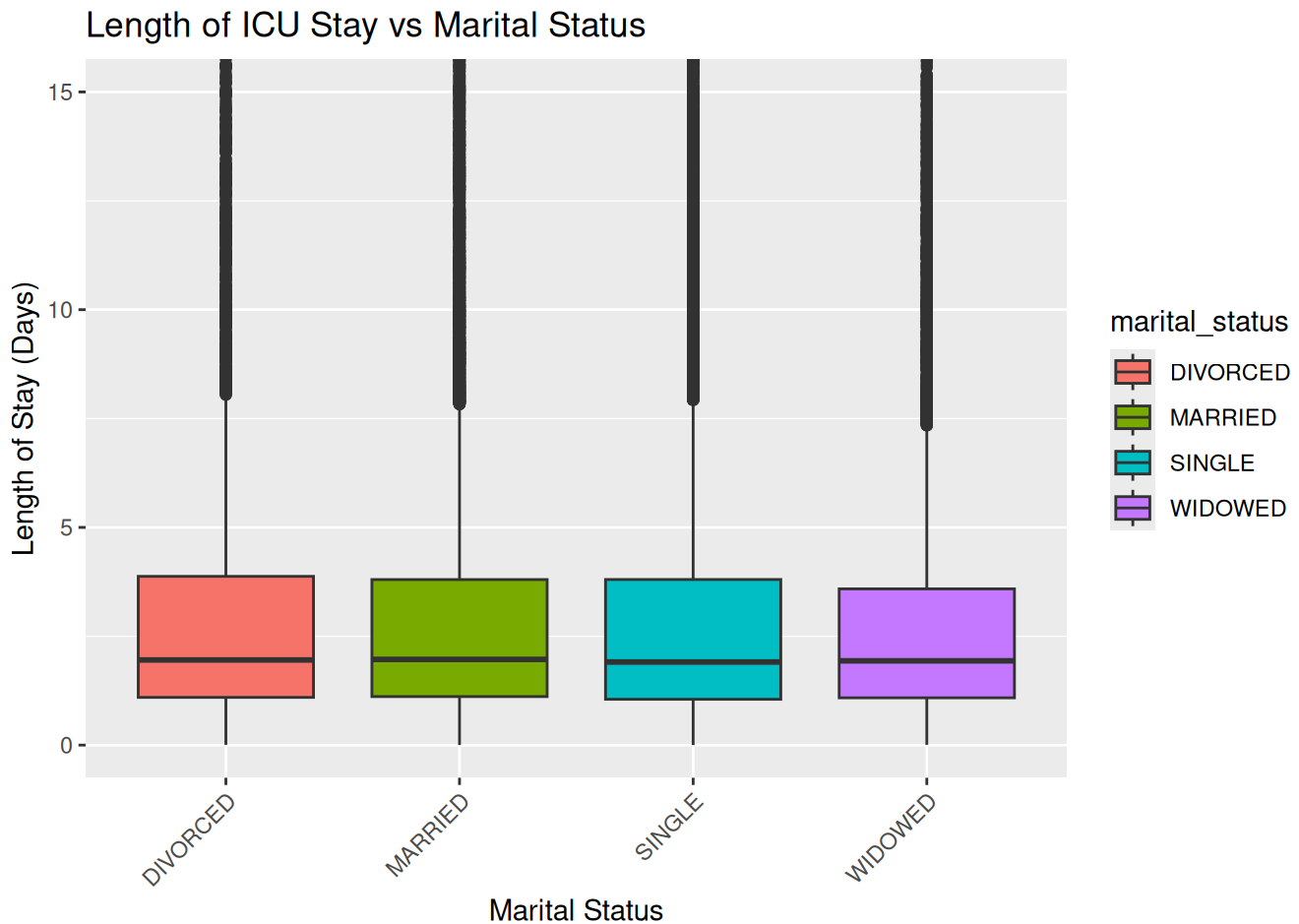


3. Boxplot of LOS versus Marital Status

The boxplot visualizes the distribution of ICU length of stay (LOS) across different marital statuses. The median LOS appears similar across all groups, indicating no significant differences in typical ICU stays between divorced, married, single, and widowed patients. Variability in LOS is present across all categories, with a few extreme outliers extending beyond 15 days. Widowed patients exhibit a slightly wider spread in LOS compared to other groups. The majority of ICU stays are short (under 5 days) across all marital statuses. The presence of outliers suggests that while most patients have short stays, a small subset experiences prolonged ICU admissions regardless of marital status. We can conclude that marital status does not strongly impact ICU length of stay, but further statistical tests are needed to confirm any potential associations. The y-axis is limited to 15 days for improved readability.

```
ggplot(mimic_icu_cohort %>% filter(!is.na(marital_status)),
       aes(x = marital_status, y = los, fill = marital_status)) +
  geom_boxplot() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Length of ICU Stay vs Marital Status",
       x = "Marital Status",
       y = "Length of Stay (Days)") +
  coord_cartesian(ylim = c(0, 15))
```

Warning: Removed 9 rows containing non-finite outside the scale range (``stat_boxplot()``).

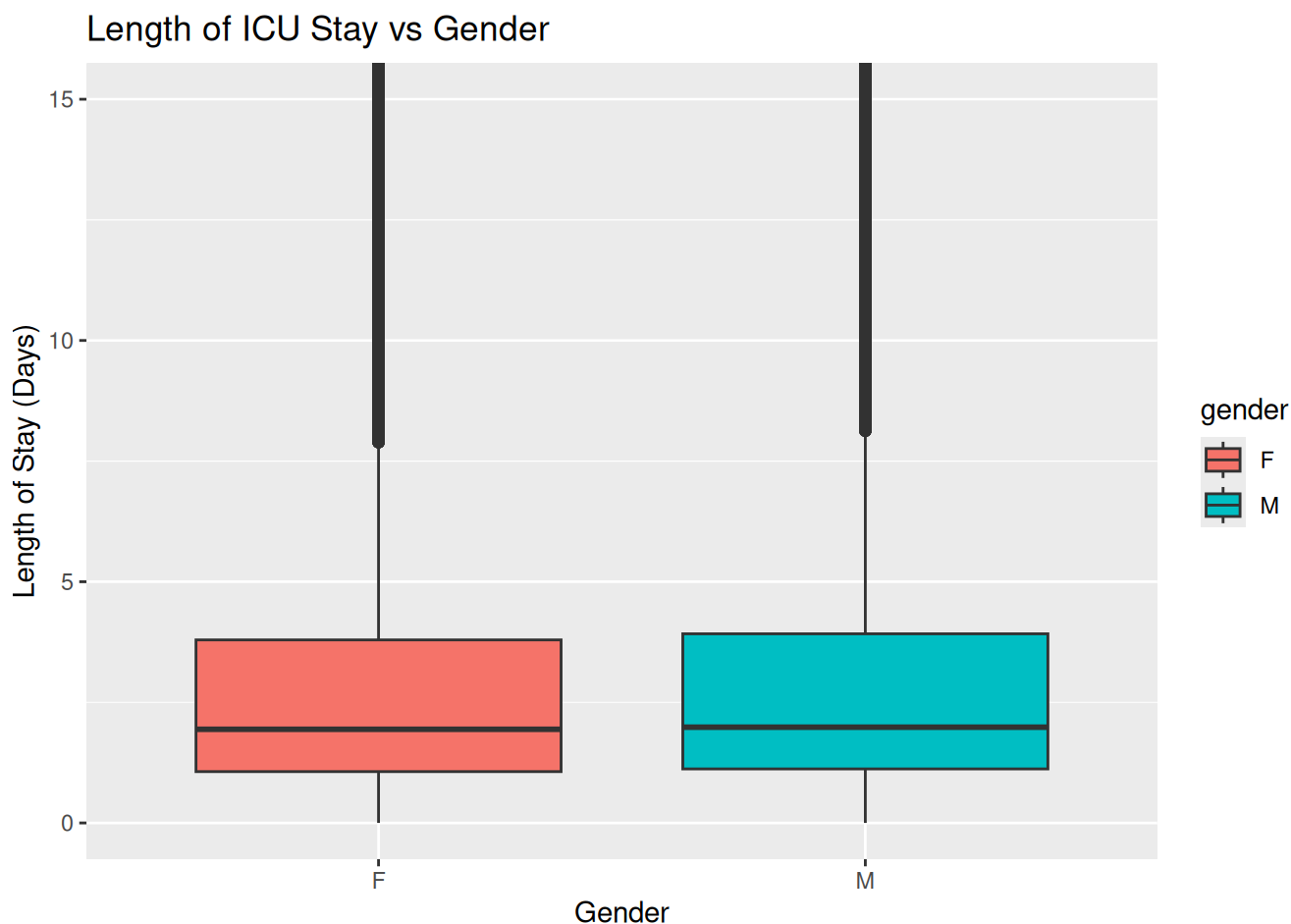


4. Boxplot of LOS versus Gender

The boxplot visualizes the distribution of ICU length of stay (LOS) by gender. The median LOS is nearly identical for both male and female patients, suggesting no substantial difference in typical ICU stays between genders. The interquartile range (IQR) is also similar, indicating that the majority of ICU stays are concentrated within a narrow range of a few days. However, both groups exhibit longer tails and extreme outliers, with some patients staying over 15 days. The presence of these outliers suggests that while most ICU stays are short, a small subset of patients experience prolonged admissions regardless of gender. The findings indicate that gender does not appear to be a strong predictor of ICU length of stay. The y-axis is limited to 15 days for better readability.

```
ggplot(mimic_icu_cohort, aes(x = gender, y = los, fill = gender)) +
  geom_boxplot() +
  labs(title = "Length of ICU Stay vs Gender",
       x = "Gender",
       y = "Length of Stay (Days)") +
  coord_cartesian(ylim = c(0, 15))
```

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



5. Boplot of LOS versus Age in Time

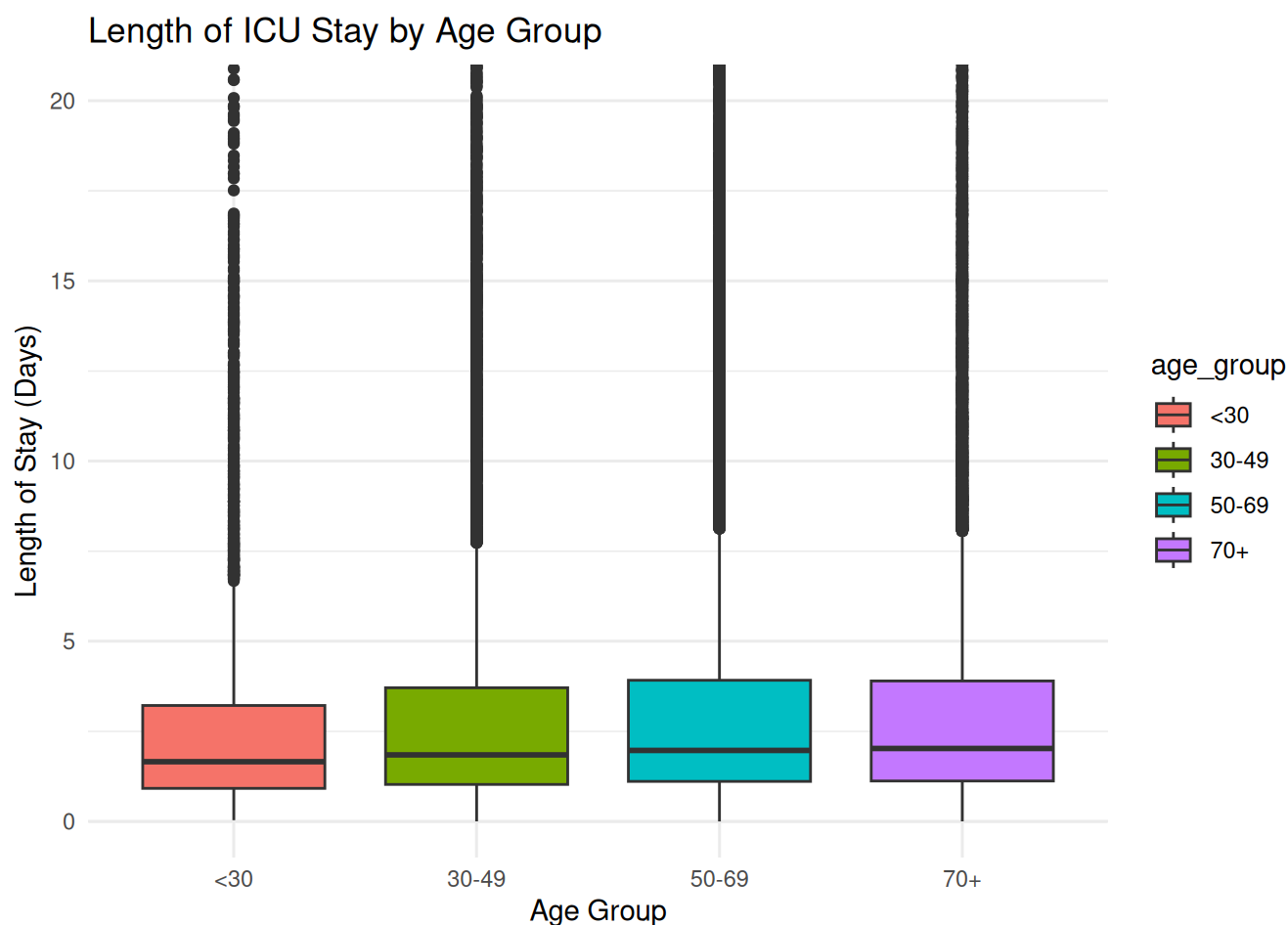
The boxplot visualizes the distribution of ICU length of stay (LOS) across different age groups (age_in_time). The median ICU stay is fairly similar across all groups, indicating that age does not have a major impact on typical ICU stays. However, older patients (70+) show a slightly wider interquartile range (IQR), suggesting increased variability in LOS. Outliers extend beyond 20 days in all age groups, showing that a small subset of patients experience prolonged ICU admissions regardless of age. The <30 age group appears to have the lowest median LOS, while older groups tend to have a slightly longer and more variable LOS. This suggests that while most ICU stays are short, older patients may have more extended stays on average, potentially due to increased medical complexity.

```
mimic_icu_cohort <- mimic_icu_cohort %>%
  mutate(age_group = case_when(
    age_at_intime < 30 ~ "<30",
    age_at_intime >= 30 & age_at_intime < 50 ~ "30-49",
    age_at_intime >= 50 & age_at_intime < 70 ~ "50-69",
    age_at_intime >= 70 ~ "70+"
  ))

ggplot(mimic_icu_cohort, aes(x = age_group, y = los, fill = age_group)) +
  geom_boxplot() +
  labs(title = "Length of ICU Stay by Age Group",
       x = "Age Group",
       y = "Length of Stay (Days)") +
```

```
coord_cartesian(ylim = c(0, 20)) +
theme_minimal()
```

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



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

The median values for all lab measurements remain consistent across different Length of ICU Stay (LOS) categories (<1 Day, 1-3 Days, 3-7 Days, >7 Days). This suggests that initial lab values upon ICU admission may not strongly correlate with LOS duration. Most lab values exhibit a stable interquartile range (IQR), indicating that patients tend to have similar initial lab values regardless of ICU stay length. However, creatinine and glucose show a slightly wider IQR, suggesting higher variability in kidney function and blood sugar levels among ICU patients. Glucose and White Blood Cell (WBC) count display higher outliers, which may indicate underlying conditions such as sepsis, infection, or metabolic disturbances. Creatinine levels have more upper outliers, possibly reflecting renal dysfunction in certain ICU patients. Electrolytes (Sodium, Potassium, Chloride, Bicarbonate) are well-distributed, with minimal outliers, suggesting relatively stable homeostasis in ICU patients. Hematocrit, a measure of red blood cell volume, remains relatively stable across LOS categories, showing no significant shifts.

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

```

mimic_icu_cohort <- mimic_icu_cohort %>%
  mutate(los_category = factor(case_when(
    los < 1 ~ "< 1 Day",
    los >= 1 & los <= 3 ~ "1-3 Days",
    los > 3 & los <= 7 ~ "3-7 Days",
    los > 7 ~ "> 7 Days"
  ), levels = c("< 1 Day", "1-3 Days", "3-7 Days", "> 7 Days"))) %>%
  filter(!is.na(los_category))

mimic_long <- mimic_icu_cohort %>%
  select(los_category, all_of(lab_vars)) %>%
  pivot_longer(
    cols = all_of(lab_vars), names_to = "Lab_Variable", values_to = "Value"
  ) %>%
  filter(!is.na(Value))

iqr_stats <- mimic_long %>%
  group_by(Lab_Variable) %>%
  summarise(
    Q1 = quantile(Value, 0.25, na.rm = TRUE),
    Q3 = quantile(Value, 0.75, na.rm = TRUE),
    IQR = Q3 - Q1,
    y_min = Q1 - 1.5 * IQR,
    y_max = Q3 + 1.5 * IQR
  )

mimic_long <- mimic_long %>%
  left_join(iqr_stats, by = "Lab_Variable") %>%
  filter(Value >= y_min & Value <= y_max)

los_colors <- c(
  "< 1 Day" = "#E74C3C",
  "1-3 Days" = "#27AE60",
  "3-7 Days" = "#F1C40F",
  "> 7 Days" = "#2980B9"
)

p <- ggplot(mimic_long, aes(
  x = los_category, y = Value, fill = los_category
)) +
  geom_boxplot(outlier.size = 1, outlier.alpha = 0.6, width = 0.6) +
  stat_summary(
    fun = median, geom = "point", size = 3, color = "black", shape = 18
  ) +
  scale_fill_manual(values = los_colors) +
  theme_minimal() +
  labs(
    title = "Boxplot of Lab Values by LOS Category",
    subtitle = "Focusing on Median, IQR, and Minor Outliers",
    x = "Length of ICU Stay (LOS) Category",

```

```

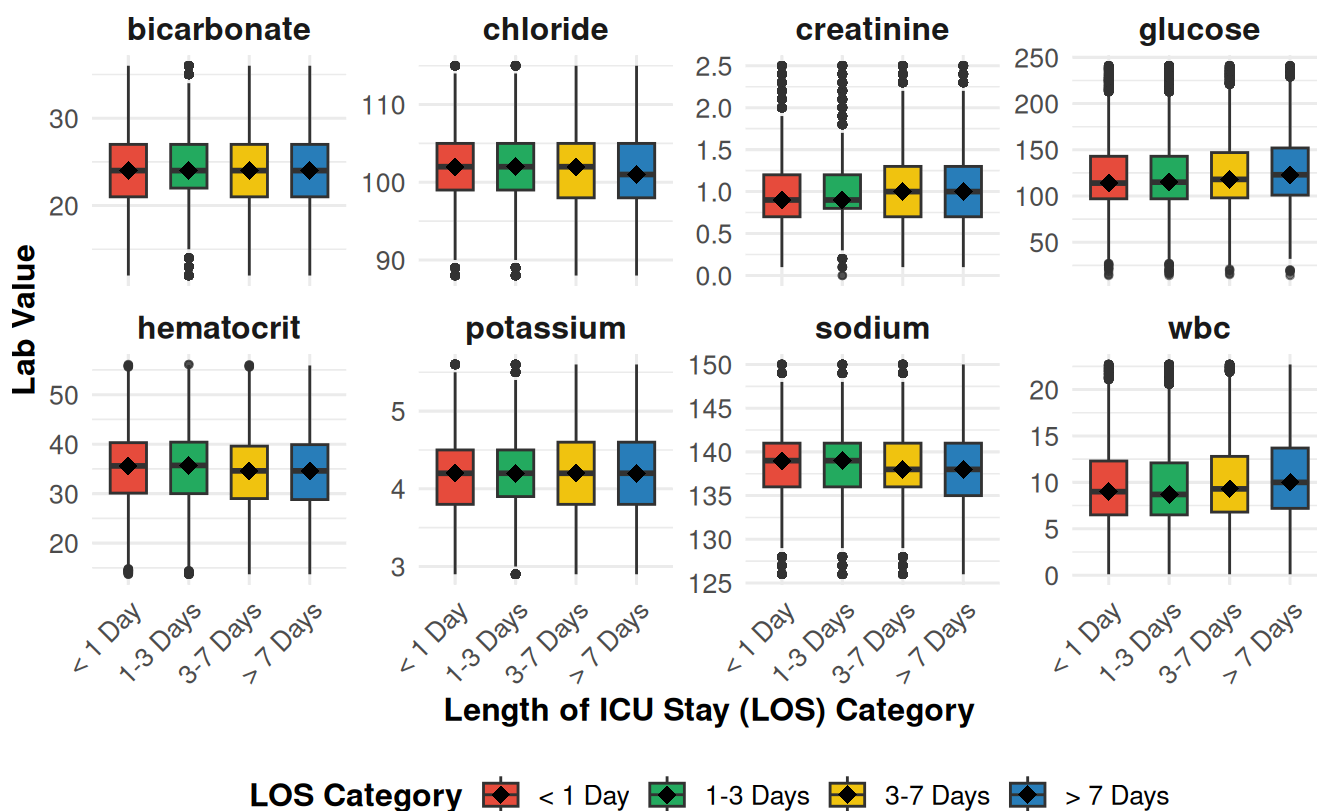
y = "Lab Value",
fill = "LOS Category"
) +
facet_wrap(~ Lab_Variable, scales = "free_y", nrow = 2) +
theme(
  legend.position = "bottom",
  legend.text = element_text(size = 10),
  legend.title = element_text(size = 12, face = "bold"),
  axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
  axis.text.y = element_text(size = 10),
  axis.title = element_text(size = 12, face = "bold"),
  plot.title = element_text(size = 16, face = "bold"),
  plot.subtitle = element_text(size = 12, face = "italic"),
  strip.text = element_text(size = 12, face = "bold")
)

print(p)

```

Boxplot of Lab Values by LOS Category

Focusing on Median, IQR, and Minor Outliers



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

The median values for each vital sign remain relatively stable across Length of ICU Stay (LOS) categories (<1 Day, 1-3 Days, 3-7 Days, >7 Days). This suggests that initial vital signs at ICU admission may not be strong predictors of ICU stay length. Heart Rate: Shows moderate variability, with higher outliers exceeding 120 bpm, but the overall distribution remains consistent across LOS groups. Non-Invasive Blood Pressure (Systolic & Diastolic): Both exhibit

a consistent range, though some extreme outliers exist at the higher end. Respiratory Rate: Displays a wider interquartile range (IQR) compared to other vital signs, but no clear trend is observed across LOS categories. Temperature (Fahrenheit): Has a tight distribution around 98-99°F, with minimal variability and a few outliers. The boxplots emphasize the interquartile range (IQR) while removing extreme outliers for better visualization. Moderate outliers are still included to provide a realistic view of variability within each vital sign.

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

mimic_icu_cohort <- mimic_icu_cohort %>%
  mutate(los_category = factor(case_when(
    los < 1 ~ "< 1 Day",
    los >= 1 & los <= 3 ~ "1-3 Days",
    los > 3 & los <= 7 ~ "3-7 Days",
    los > 7 ~ "> 7 Days"
  ), levels = c("< 1 Day", "1-3 Days", "3-7 Days", "> 7 Days")))

mimic_long <- mimic_icu_cohort %>%
  select(los_category, all_of(vital_vars)) %>%
  pivot_longer(
    cols = all_of(vital_vars), names_to = "Vital_Sign", values_to = "Value"
  )

iqr_stats <- mimic_long %>%
  group_by(Vital_Sign) %>%
  summarise(
    Q1 = quantile(Value, 0.25, na.rm = TRUE),
    Q3 = quantile(Value, 0.75, na.rm = TRUE),
    IQR = Q3 - Q1,
    y_min = Q1 - 1.5 * IQR,
    y_max = Q3 + 1.5 * IQR
  )

mimic_long <- mimic_long %>%
  left_join(iqr_stats, by = "Vital_Sign") %>%
  filter(Value >= y_min & Value <= y_max)

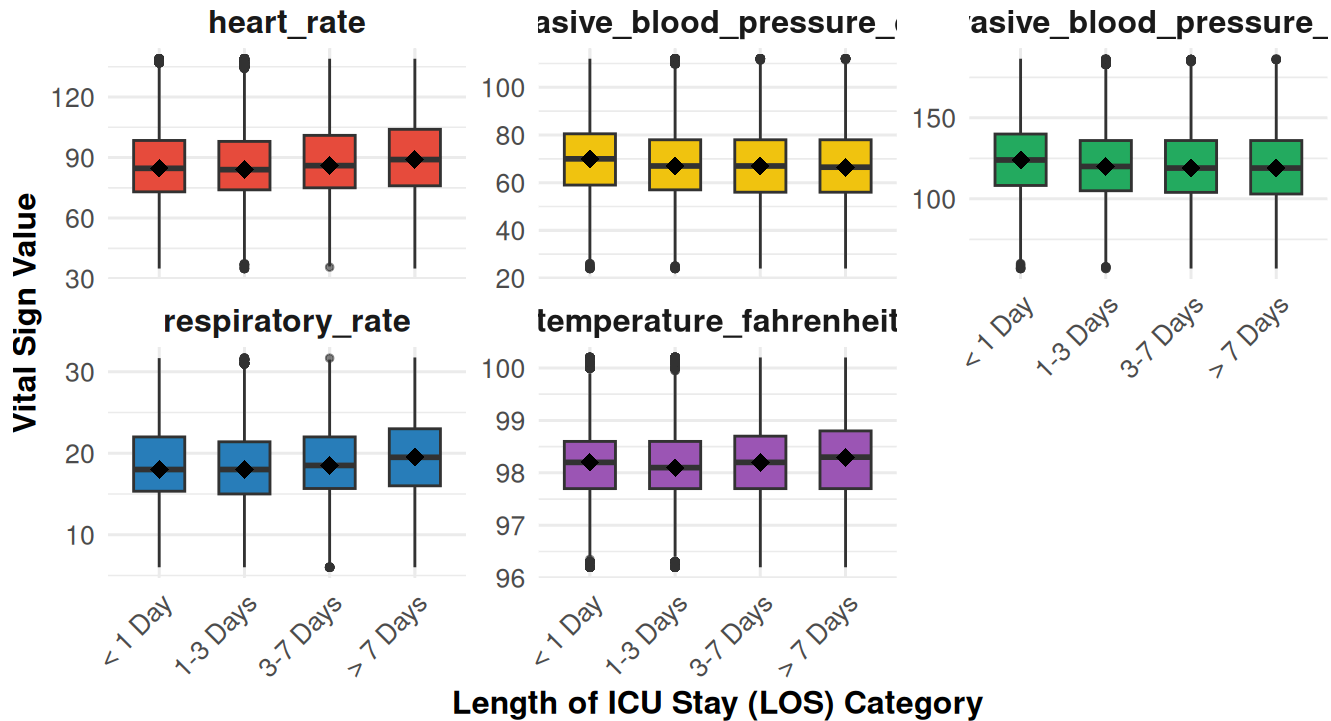
vital_colors <- c(
  "heart_rate" = "#E74C3C",
  "non_invasive_blood_pressure_systolic" = "#27AE60",
  "non_invasive_blood_pressure_diastolic" = "#F1C40F",
  "respiratory_rate" = "#2980B9",
  "temperature_fahrenheit" = "#9B59B6"
)

p <- ggplot(mimic_long, aes(
  x = los_category, y = Value, fill = Vital_Sign
```

```
)) +  
geom_boxplot(outlier.size = 1, outlier.alpha = 0.6, width = 0.6) +  
stat_summary(  
  fun = median, geom = "point", size = 3, color = "black", shape = 18  
) +  
scale_fill_manual(values = vital_colors) + # Set custom colors  
theme_minimal() +  
labs(  
  title = "Boxplot of Vital Signs by LOS Category",  
  subtitle = "Focusing on Median, IQR, and Moderate Outliers",  
  x = "Length of ICU Stay (LOS) Category",  
  y = "Vital Sign Value",  
  fill = "Vital Sign"  
) +  
facet_wrap(~ Vital_Sign, scales = "free_y", nrow = 2) + # Arrange in 2 rows  
theme(  
  legend.position = "bottom",  
  legend.text = element_text(size = 10),  
  legend.title = element_text(size = 12, face = "bold"),  
  axis.text.x = element_text(angle = 45, hjust = 1, size = 10),  
  axis.text.y = element_text(size = 10),  
  axis.title = element_text(size = 12, face = "bold"),  
  plot.title = element_text(size = 16, face = "bold"),  
  plot.subtitle = element_text(size = 12, face = "italic"),  
  strip.text = element_text(size = 12, face = "bold")  
)  
  
print(p)
```

Boxplot of Vital Signs by LOS Category

Focusing on Median, IQR, and Moderate Outliers



rate non_invasive_blood_pressure_diastolic non_invasive_blood_pressure_systolic respiratory_r

Length of ICU stay los vs first ICU unit

The median LOS varies significantly across different ICU Units. Surgical and specialized ICUs, such as Surgery/Vascular/Intermediate, Neurology, and Medicine ICUs tend to have longer median LOS, indicating that patients in these units require extended care or recovery periods. In contrast, general ICUs and step down units, including Med/Surg, MICU/SICU, Trauma SICU, and PACU exhibit shorter median LOS, likely due to lower severity cases. Among the units with the longest ICU stays, Surgery/Vascular Intermediate and Neurology ICUs stand out with higher median LOS and wider interquartile ranges, reflecting the complexity of cases managed in these units. Similarly, Surgery/Trauma and Intensive Care Units (ICU) also display prolonged hospitalizations, suggesting severe cases requiring extended treatment. The distribution of LOS reveals a high concentration of outliers in most ICU units, with some extreme LOS values exceeding 30+ days. These extended stays are likely associated with complicated medical conditions or post surgery treatments. The findings suggest that ICU type significantly impacts hospitalization duration, and we are able to see from the results of the Kruskal-Wallis test that there is statistically significant differences in LOS among ICU units with a p-value of less than 0.05.

```
mimic_icu_cohort$first_careunit <- str_wrap(
  mimic_icu_cohort$first_careunit, width = 15
)
icu_unit_summary <- mimic_icu_cohort %>%
  group_by(first_careunit) %>%
  summarise(
    los_mean = mean(los, na.rm = TRUE),
    los_median = median(los, na.rm = TRUE),
```



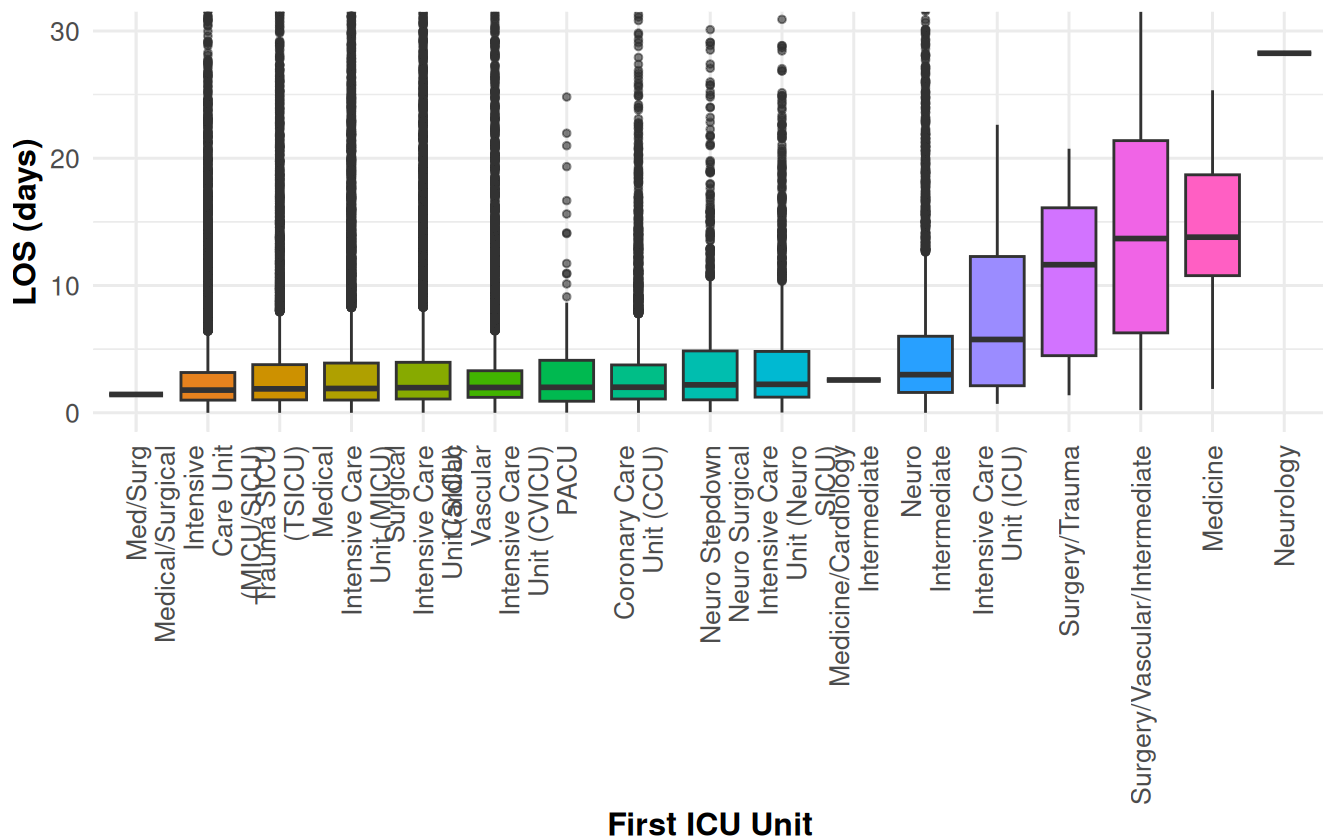
```
    los_sd = sd(los, na.rm = TRUE),
    los_min = min(los, na.rm = TRUE),
    los_max = max(los, na.rm = TRUE),
    count = n()
  ) %>%
  arrange(desc(los_median))

mimic_icu_cohort <- mimic_icu_cohort %>%
  mutate(first_careunit = reorder(
    first_careunit, los, FUN = median, na.rm = TRUE
  ))

ggplot(mimic_icu_cohort, aes(
  x = first_careunit, y = los, fill = first_careunit
)) +
  geom_boxplot(outlier.size = 1, outlier.alpha = 0.6) +
  coord_cartesian(ylim = c(0, 30)) +
  theme_minimal() +
  labs(
    title = "Length of ICU Stay vs First ICU Unit",
    subtitle = "ICU Units Ordered by Median LOS",
    x = "First ICU Unit",
    y = "LOS (days)"
  ) +
  theme(
    legend.position = "none",
    axis.text.x = element_text(
      angle = 90, hjust = 1, vjust = 0.5, size = 10
    ),
    axis.text.y = element_text(size = 10),
    axis.title = element_text(size = 12, face = "bold"),
    plot.title = element_text(size = 16, face = "bold"),
    plot.subtitle = element_text(size = 12, face = "italic")
  )
)
```

Length of ICU Stay vs First ICU Unit

ICU Units Ordered by Median LOS



```
# Perform Kruskal-Wallis test
kruskal_test <- kruskal.test(los ~ first_careunit, data = mimic_icu_cohort)
print(kruskal_test)
```

Kruskal-Wallis rank sum test

data: los by first_careunit

Kruskal-Wallis chi-squared = 1464.8, df = 16, p-value < 2.2e-16

End of Homework #3