Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

Kiana Mohammadinik and 205928003

Table of contents

Display machine information for reproducibility:

```
sessionInfo()
```

```
R version 4.3.0 (2023-04-21)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS 14.4.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
```

```
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-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_4.3.0    fastmap_1.1.1     cli_3.6.3         tools_4.3.0
 [5] htmltools_0.5.8.1 rstudioapi_0.14   yaml_2.3.8        rmarkdown_2.29
 [9] knitr_1.45        jsonlite_1.8.8    xfun_0.50         digest_0.6.34
[13] rlang_1.1.4       evaluate_0.23
```

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

```
library(arrow)
```

```
Warning: package 'arrow' was built under R version 4.3.3
```

```

Attaching package: 'arrow'

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

    timestamp
```

```
library(gtsummary)
```

```
Warning: package 'gtsummary' was built under R version 4.3.3
```

```
library(memuse)
```

```
Warning: package 'memuse' was built under R version 4.3.3
```

```
library(pryr)
```

```
Attaching package: 'pryr'

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

    where
```

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

```
Warning: package 'R.utils' was built under R version 4.3.1

Loading required package: R.oo

Warning: package 'R.oo' was built under R version 4.3.1

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 v1.26.0 (2024-01-24 05:12:50 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, warnings
```

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.1

Warning: package 'tidyr' was built under R version 4.3.1

Warning: package 'dplyr' was built under R version 4.3.1

Warning: package 'stringr' was built under R version 4.3.1

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.1
v purrr     1.0.1
```

```
-- 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
```

```
library(ggplot2)
library(dplyr)
library(lubridate)
library(stringr)
library(duckdb)
```

```
Warning: package 'duckdb' was built under R version 4.3.3
```

```
Loading required package: DBI
```

```
library(tidyr)
```

Display your machine memory.

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

```
Totalram:  16.000 GiB
Freeram:    5.032 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.

0.1 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.

0.1.1 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:

```
patients <- 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.
```

```
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
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")
```

```
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.
```

```
labevents_pq <- read_parquet("labevents_pq/part-0.parquet")
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.
```

```
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.
```

```
d_icd_procedures <- 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.
```

```
d_icd_diagnoses <- 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.
```

```
icustays <- 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.
```

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

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

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

```r
chartevents_pq <- read_parquet("chartevents_filtered.parquet")
```

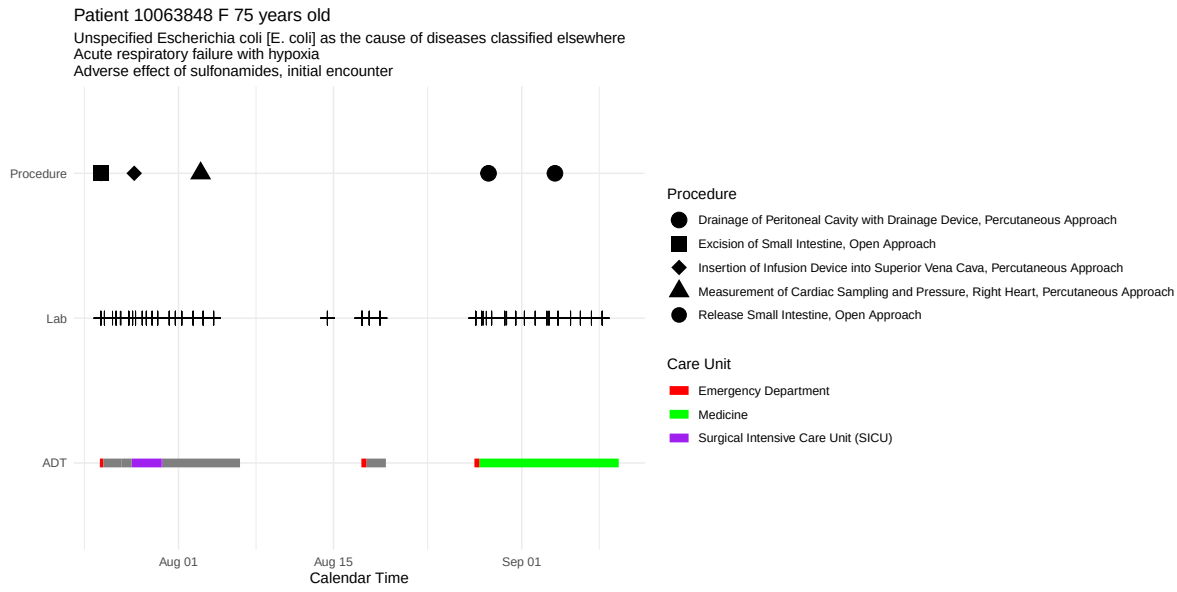
```r
# Filter data for patient 10063848
subject_id <- 10063848
patient_info <- patients %>% filter(subject_id == !!subject_id)
admissions_info <- admissions %>% filter(subject_id == !!subject_id)
transfers_info <- transfers %>% filter(subject_id == !!subject_id)
labevents_info <- labevents_pq %>% filter(subject_id == !!subject_id)
procedures_info <- procedures %>% filter(subject_id == !!subject_id)
diagnoses_info <- diagnoses %>% filter(subject_id == !!subject_id)
```

0.2 plot

```r
diagnoses_info <- diagnoses_info %>%
  mutate(icd_code = str_pad(icd_code, width = 5, pad = "0"))

diagnoses_info <- diagnoses_info %>%
  left_join(d_icd_diagnoses, by = c("icd_code", "icd_version"))

long_title_col <- grep("long_title", colnames(diagnoses_info), value = TRUE)

if ("long_title" %in% long_title_col) {
  diagnoses_info <- diagnoses_info %>% rename(diagnosis_name = long_title)
} else if ("long_title.x" %in% long_title_col) {
  diagnoses_info <- diagnoses_info %>% rename(diagnosis_name = long_title.x)
} else if ("long_title.y" %in% long_title_col) {
  diagnoses_info <- diagnoses_info %>% rename(diagnosis_name = long_title.y)
} else {
  stop("long_title column not found in diagnoses_info")
}

top_diagnoses <- diagnoses_info %>%
  filter(!is.na(diagnosis_name)) %>%
  count(diagnosis_name, sort = TRUE) %>%
  head(3) %>%
  pull(diagnosis_name)

top_diagnoses_text <- ifelse(length(top_diagnoses) > 0,
  paste(top_diagnoses, collapse = "\n"), "")
```

```r
patient_summary <- paste(
  "Patient",
  subject_id,
  ifelse(is.na(patient_info$gender), "", patient_info$gender),
  ifelse(is.na(patient_info$anchor_age), "",
  paste(patient_info$anchor_age, "years old"))
) %>% str_squish()

transfers_info <- transfers_info %>%
  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))

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

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

procedure_name_cols <- grep("long_title", colnames(procedures_info),
  value = TRUE)

if (length(procedure_name_cols) > 1) {
  procedures_info <- procedures_info %>%
  select(-one_of(procedure_name_cols[-1])) %>%
  rename(procedure_name = procedure_name_cols[1])
} else if (length(procedure_name_cols) == 1) {
  procedures_info <- procedures_info %>%
  rename(procedure_name = procedure_name_cols[1])
} else {
  stop("procedure_name column not found in procedures_info")
}

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

care_unit_colors <- c("Emergency Department" = "red",
                      "Medicine" = "green",
                      "Neurology" = "cyan",
                      "Surgical Intensive Care Unit (SICU)" = "purple")

procedure_shapes <- setNames(seq(15, 15 +
```

```
  length(unique(procedures_info$procedure_name)) - 1),
  unique(procedures_info$procedure_name))

plot <- ggplot() +
  geom_segment(data = transfers_info,
  aes(x = intime, xend = outtime,
  y = "ADT", yend = "ADT", color = careunit),
  linewidth = 3) +
  geom_point(data = labevents_info,
  aes(x = chartdate, y = "Lab"), shape = 3, size = 3) +
  geom_point(data = procedures_info,
  aes(x = chartdate, y = "Procedure", shape = procedure_name), size = 5) +
  scale_color_manual(values = care_unit_colors) +
  scale_shape_manual(values = procedure_shapes, drop = FALSE) +
  theme_minimal() +
  labs(title = patient_summary,
  subtitle = top_diagnoses_text,
  x = "Calendar Time",
  y = NULL,
  color = "Care Unit",
  shape = "Procedure")
print(plot)
```



0.2.1 Q1.2 ICU stays

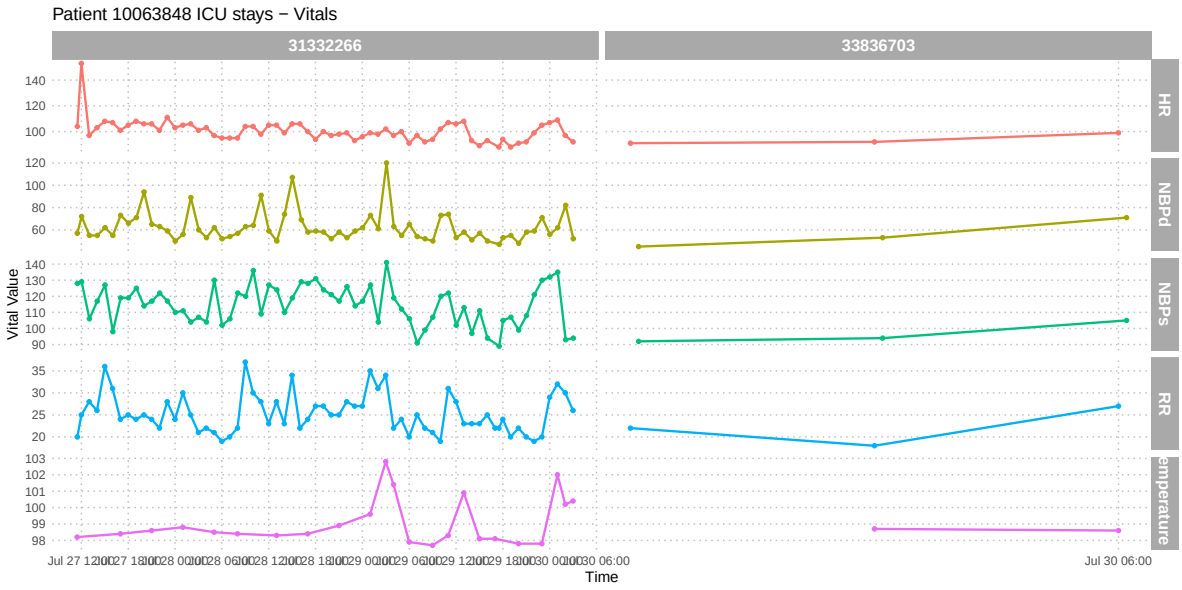
Solution:

```
subject_id_of_interest <- 10063848
subject_stays <- icustays %>%
  filter(subject_id == subject_id_of_interest) %>%
  select(stay_id, intime, outtime)
chartevents_filtered <- chartevents_pq %>%
  filter(subject_id == subject_id_of_interest) %>%
  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")
chartevents_with_labels <- chartevents_with_labels %>%
  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.
```



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`.

0.3 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
```

0.3.1 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.
```

0.3.2 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: The number of unique values in `subject_id`

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

cat("Number of unique subjects:", num_unique_subjects$n, "\n")
```

```
Number of unique subjects: 65366
```

Checking if a subject has had multiple ICU stays

```
icu_stay_counts <- icustays_tble %>%
  group_by(subject_id) %>%
  summarize(num_stays = n(), .groups = "drop")
cat("Max ICU stays by a single patient:", max(icu_stay_counts$num_stays), "\n")
```

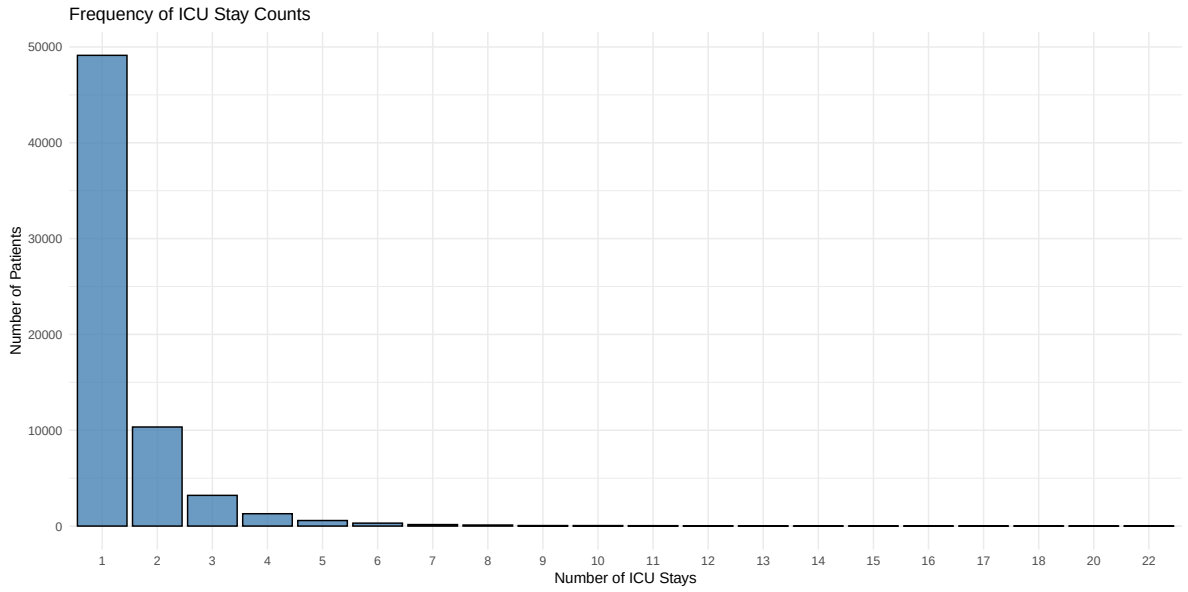
```
Max ICU stays by a single patient: 41
```

```
num_multiple_stays <- icu_stay_counts %>%
  filter(num_stays > 1) %>%
  count()
cat("Number of patients with multiple ICU stays:", num_multiple_stays$n, "\n")
```

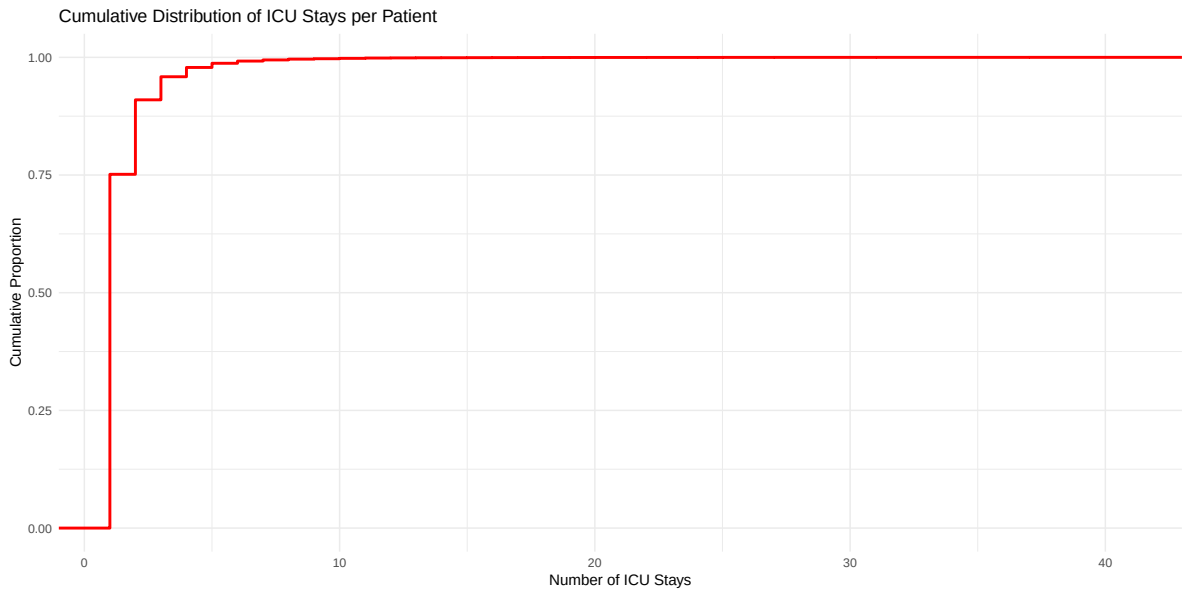
```
Number of patients with multiple ICU stays: 16242
```

Summarizeing the number of ICU stays per `subject_id` by graphs

```
# Bar Chart of ICU Stay Counts
icu_stay_counts %>%
  count(num_stays) %>%
  arrange(desc(n)) %>%
  head(20) %>%
  ggplot(aes(x = factor(num_stays), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black"
           , alpha = 0.8) +
  labs(
    title = "Frequency of ICU Stay Counts",
    x = "Number of ICU Stays",
    y = "Number of Patients"
  ) +
  theme_minimal()
```




```r
# Cumulative Frequency Plot
ggplot(icu_stay_counts, aes(x = num_stays)) +
  stat_ecdf(geom = "step", color = "red", size = 1) +
  labs(
    title = "Cumulative Distribution of ICU Stays per Patient",
    x = "Number of ICU Stays",
    y = "Cumulative Proportion"
  ) +
  theme_minimal()
```



0.4 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,HOM
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,HOM
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
```

0.4.1 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.
```

0.4.2 Q3.2 Summary and visualization

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

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

According to the MIMIC-IV documentation,

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

Solution: Number of admissions per patient

```
admissions_tble <- admissions_tble %>%
  mutate(
    admission_hour = hour(admittime),
    admission_minute = minute(admittime),
    los_days = as.numeric(difftime(dischtime, admittime, units = "days"))
  )
```

```r
admissions_tble <- admissions_tble %>%
  mutate(
    admission_hour = hour(admittime),
    admission_minute = minute(admittime),
    los_days = as.numeric(difftime(dischtime, admittime, units = "days"))
  )


summary(admissions_tble$admission_hour)
```

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

```r
summary(admissions_tble$admission_minute)
```

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

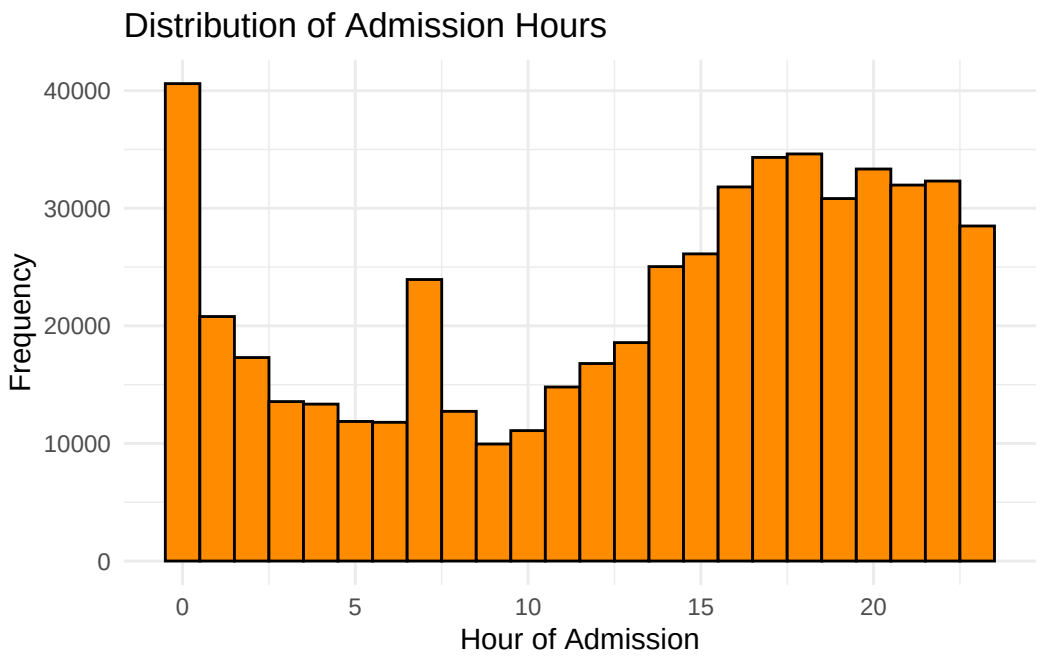
```r
summary(admissions_tble$los_days)
```

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

Explanation of Patterns:
Admission Hour: There is a peak around 0 AM and 7 AM, which is unusual
0 AM might reflect end-of-day admissions or data recording, while 7 AM may be related to shift changes.
Admission Minute: The 15-minute peak intervals (e.g., 15, 30, 45) might be due to rounding in documentation.
Length of Stay: Most patients stay between 1 and 5 days, with outliers staying longer due to severe conditions.

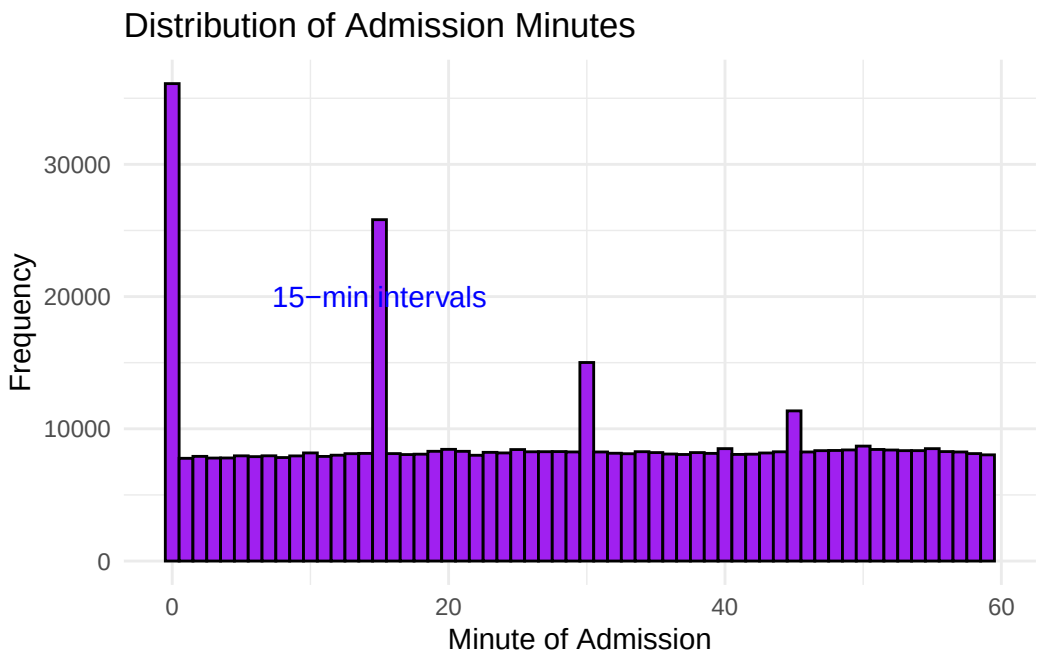
Admission hour

```r
ggplot(admissions_tble, aes(x = admission_hour)) +
  geom_histogram(binwidth = 1, fill = "darkorange", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Admission Hours", x = "Hour of Admission", y = "Frequency")
```



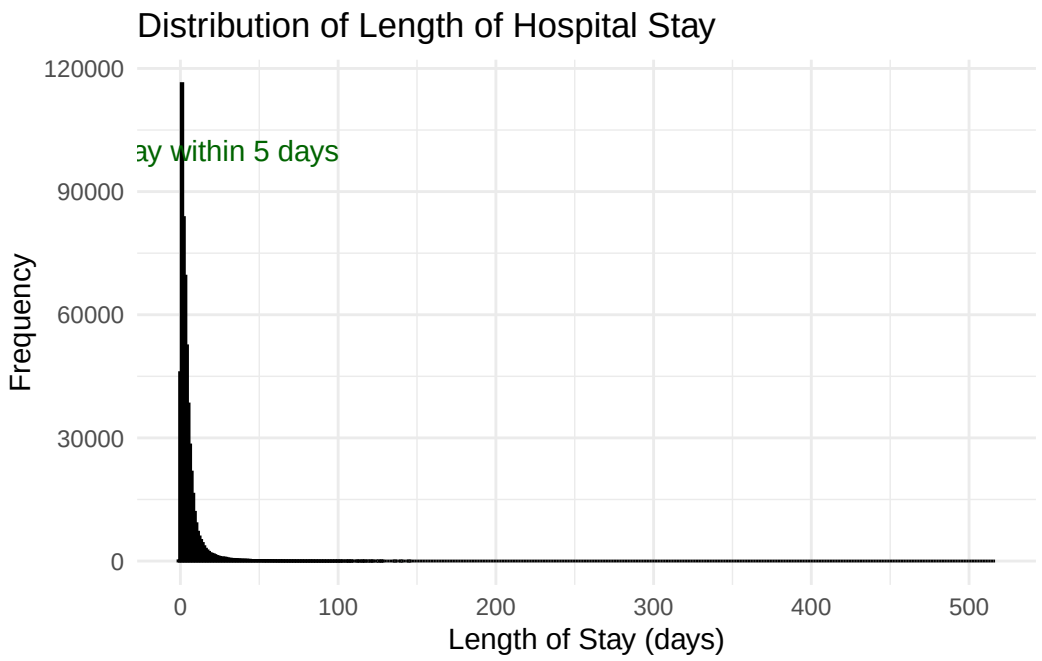
Admission minute

```
ggplot(admissions_tble, aes(x = admission_minute)) +
  geom_histogram(binwidth = 1, fill = "purple", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Admission Minutes", x = "Minute of Admission", y = "Frequency
  annotate("text", x = 15, y = 20000, label = "15-min intervals", color = "blue")
```



Length of hospital stay (from admission to discharge)

```
ggplot(admissions_tble, aes(x = los_days)) +
  geom_histogram(binwidth = 1, fill = "seagreen", color = "black") +
  theme_minimal() +
  labs(title = "Distribution of Length of Hospital Stay", x = "Length of Stay (days)", y = "
  annotate("text", x = 5, y = 100000, label = "Most stay within 5 days", color = "darkgreen"
```



0.5 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,
```

0.5.1 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.
```

```
glimpse(patients_tble)
```

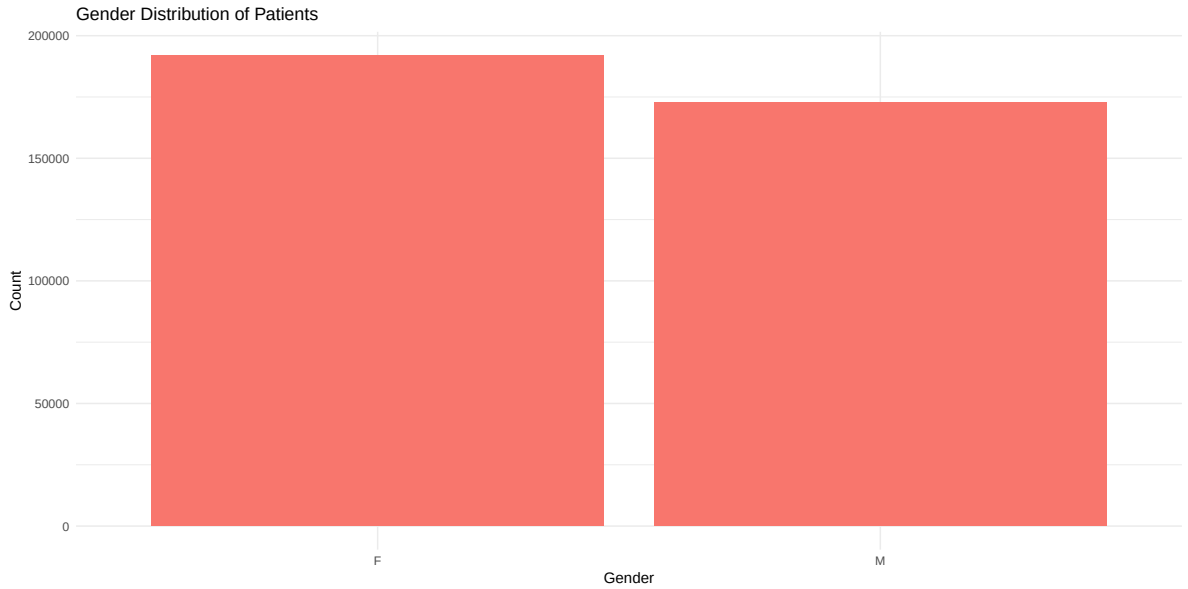
```
Rows: 364,627
Columns: 6
$ subject_id        <dbl> 10000032, 10000048, 10000058, 10000068, 10000084, 10~
$ gender            <chr> "F", "F", "F", "F", "M", "F", "M", "M", "F", "M", "F~
$ anchor_age        <dbl> 52, 23, 33, 19, 72, 27, 25, 24, 48, 60, 59, 34, 20, ~
$ anchor_year       <dbl> 2180, 2126, 2168, 2160, 2160, 2136, 2163, 2154, 2174~
$ anchor_year_group <chr> "2014 - 2016", "2008 - 2010", "2020 - 2022", "2008 -~
$ dod               <date> 2180-09-09, NA, NA, NA, 2161-02-13, NA, NA, NA, NA,~
```

0.5.2 Q4.2 Summary and visualization

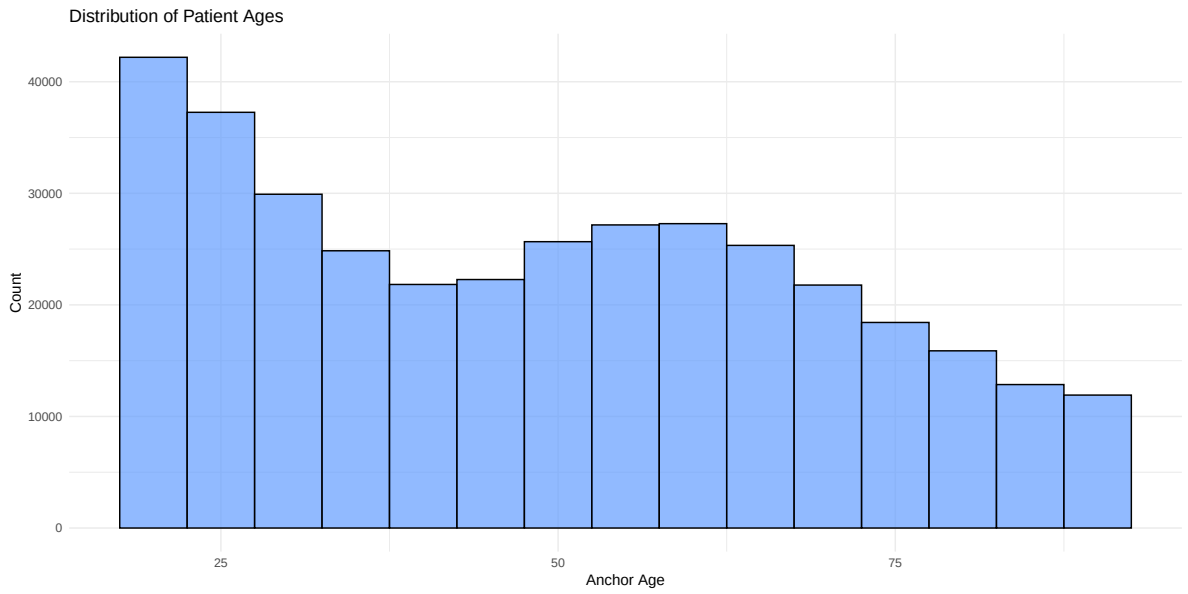
Summarize variables `gender` and `anchor_age` by graphics, and explain any patterns you see.

Solution:

```
# Gender Distribution
ggplot(patients_tble, aes(x = gender)) +
  geom_bar(fill = "#F8766D") +
  labs(title = "Gender Distribution of Patients", x = "Gender", y = "Count") +
  theme_minimal()
```




```r
# Anchor Age Distribution
ggplot(patients_tble, aes(x = anchor_age)) +
  geom_histogram(binwidth = 5, fill = "#619CFF", color = "black", alpha = 0.7) +
  labs(title = "Distribution of Patient Ages", x = "Anchor Age", y = "Count") +
  theme_minimal()
```



Explanation of Gender Distribution:
The dataset has a fairly balanced gender distribution, with slightly more females than males.

Explanation of Anchor Age Distribution:
The age distribution shows that most patients are between 50 and 80 years old, which is expected given the ICU setting.
The sharp peak at age 91 is due to the MIMIC-IV data de-identification process, where patients older than 89 are all assigned an age of 91.

0.6 Q5. Lab results

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

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

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

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

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

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

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

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:

```
d_labitems_tble <- read_csv("~/mimic/hosp/d_labitems.csv.gz") %>%
  filter(itemid %in% c(50912, 50971, 50983, 50902, 50882, 51221, 51301, 50931))
```

```
Rows: 1650 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): label, fluid, category
dbl (1): itemid

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_tble <- open_dataset(sources = "labevents_pq/part-0.parquet", format = "parquet")
  to_duckdb() |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in% local(d_labitems_tble$itemid)) |>
  left_join(select(icustays_tble, subject_id, stay_id, intime), by = c("subject_id" = "subje
  filter(storetime < intime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_max(storetime, n = 1) |>
  select(-storetime, -intime) |>
  ungroup() |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  rename_with(~ str_to_lower(d_labitems_tble$label[match(.x, as.character(d_labitems_tble$ite
  collect() |>
  arrange(subject_id, stay_id) |>
  relocate(subject_id, stay_id, .before = everything())
cat("Final number of rows:", nrow(labevents_tble), "\n")
```

```
Final number of rows: 88086
```

0.7 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.

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:

```
vital_ids <- c(
  "heart_rate" = 220045,
  "non_invasive_blood_pressure_systolic" = 220179,
  "non_invasive_blood_pressure_diastolic" = 220180,
  "temperature_fahrenheit" = 223761,
  "respiratory_rate" = 220210
)

subset_chartevents <- chartevents_pq %>%
  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 u
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.
```

```r
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_"
  ) %>%
  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)

cat("Final number of rows:", nrow(chartevents_tble), "\n")
```

```
Final number of rows: 94363
```

0.8 Q7. Putting things together

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

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

- the first vital measurements during the ICU stay in `chartevents_tble`

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

Solution:

```
icustays_age <- icustays_tble %>%
  mutate(intime_year = lubridate::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_filtered <- icustays_tble %>%
  left_join(age_at_intime, by = c("subject_id", "stay_id")) %>%
  inner_join(patients_tble %>%
               select(subject_id, anchor_age, anchor_year, anchor_year_group, dod, gender),
             by = "subject_id") %>%
  filter(age_at_intime >= 18)

admissions_selected <- admissions_tble

vitals_selected <- chartevents_tble
labs_selected <- labevents_tble

mimic_icu_cohort <- icustays_filtered %>%
  left_join(admissions_selected, by = c("subject_id", "hadm_id")) %>%
  left_join(vitals_selected, by = c("subject_id", "stay_id")) %>%
  left_join(labs_selected, by = c("subject_id", "stay_id")) %>%
  distinct() %>%
  arrange(subject_id, hadm_id, stay_id)

print(mimic_icu_cohort)
```

```
# A tibble: 94,458 x 44
   subject_id  hadm_id  stay_id first_careunit last_careunit intime
        <dbl>    <dbl>    <dbl> <chr>          <chr>         <dttm>
 1   10000032 29079034 39553978 Medical Inten~ Medical Inte~ 2180-07-23 14:00:00
 2   10000690 25860671 37081114 Medical Inten~ Medical Inte~ 2150-11-02 19:37:00
 3   10000980 26913865 39765666 Medical Inten~ Medical Inte~ 2189-06-27 08:42:00
 4   10001217 24597018 37067082 Surgical Inte~ Surgical Int~ 2157-11-20 19:18:02
```

```
 5   10001217 27703517 34592300 Surgical Inte~ Surgical Int~ 2157-12-19 15:42:24
 6   10001725 25563031 31205490 Medical/Surgi~ Medical/Surg~ 2110-04-11 15:52:22
 7   10001843 26133978 39698942 Medical/Surgi~ Medical/Surg~ 2134-12-05 18:50:03
 8   10001884 26184834 37510196 Medical Inten~ Medical Inte~ 2131-01-11 04:20:05
 9   10002013 23581541 39060235 Cardiac Vascu~ Cardiac Vasc~ 2160-05-18 10:00:53
10   10002114 27793700 34672098 Coronary Care~ Coronary Car~ 2162-02-17 23:30:00
# i 94,448 more rows
# i 38 more variables: outtime <dttm>, los <dbl>, age_at_intime <dbl>,
#   anchor_age <dbl>, anchor_year <dbl>, anchor_year_group <chr>, dod <date>,
#   gender <chr>, admittime <dttm>, dischtime <dttm>, deathtime <dttm>,
#   admission_type <chr>, admit_provider_id <chr>, admission_location <chr>,
#   discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dttm>, edouttime <dttm>, ...
```

```
colnames(mimic_icu_cohort)
```

```
 [1] "subject_id"
 [2] "hadm_id"
 [3] "stay_id"
 [4] "first_careunit"
 [5] "last_careunit"
 [6] "intime"
 [7] "outtime"
 [8] "los"
 [9] "age_at_intime"
[10] "anchor_age"
[11] "anchor_year"
[12] "anchor_year_group"
[13] "dod"
[14] "gender"
[15] "admittime"
[16] "dischtime"
[17] "deathtime"
[18] "admission_type"
[19] "admit_provider_id"
[20] "admission_location"
[21] "discharge_location"
[22] "insurance"
[23] "language"
[24] "marital_status"
[25] "race"
[26] "edregtime"
```

```
[27] "edouttime"
[28] "hospital_expire_flag"
[29] "admission_hour"
[30] "admission_minute"
[31] "los_days"
[32] "heart_rate"
[33] "non_invasive_blood_pressure_systolic"
[34] "non_invasive_blood_pressure_diastolic"
[35] "respiratory_rate"
[36] "temperature_fahrenheit"
[37] "potassium"
[38] "chloride"
[39] "bicarbonate"
[40] "glucose"
[41] "hematocrit"
[42] "white blood cells"
[43] "creatinine"
[44] "sodium"
```

0.9 Q8. Exploratory data analysis (EDA)

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

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

```
mimic_icu_cohort %>%
  select(los, race, insurance, marital_status, gender, age_at_intime) %>%
  summary()
```

```
      los                race              insurance         marital_status
 Min.   :  0.00125   Length:94458       Length:94458       Length:94458
 1st Qu.:  1.09621   Class :character   Class :character   Class :character
 Median :  1.96565   Mode  :character   Mode  :character   Mode  :character
 Mean   :  3.63002
 3rd Qu.:  3.86258
 Max.   :226.40308
 NA's   :14
    gender          age_at_intime
 Length:94458       Min.   : 18.00
 Class :character   1st Qu.: 55.00
```

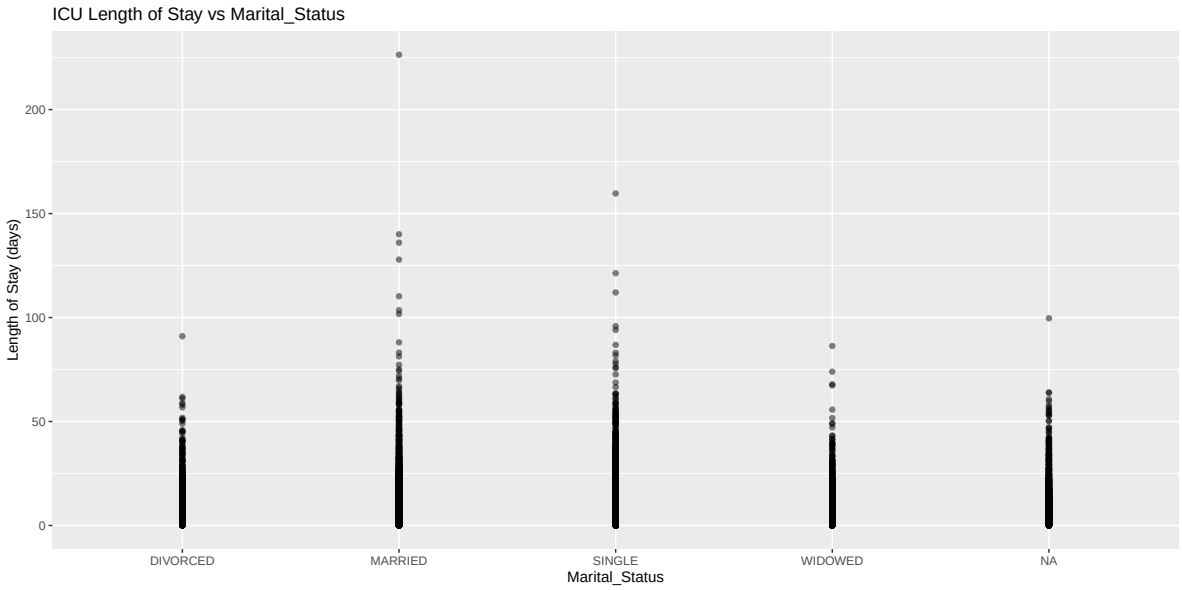
```
 Mode   :character   Median : 66.00
                     Mean   : 64.79
                     3rd Qu.: 77.00
                     Max.   :103.00
```

```
ggplot(mimic_icu_cohort, aes(x = marital_status, y = los)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
  labs(title = "ICU Length of Stay vs Marital_Status", x = "Marital_Status",
       y = "Length of Stay (days)")
```

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

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

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

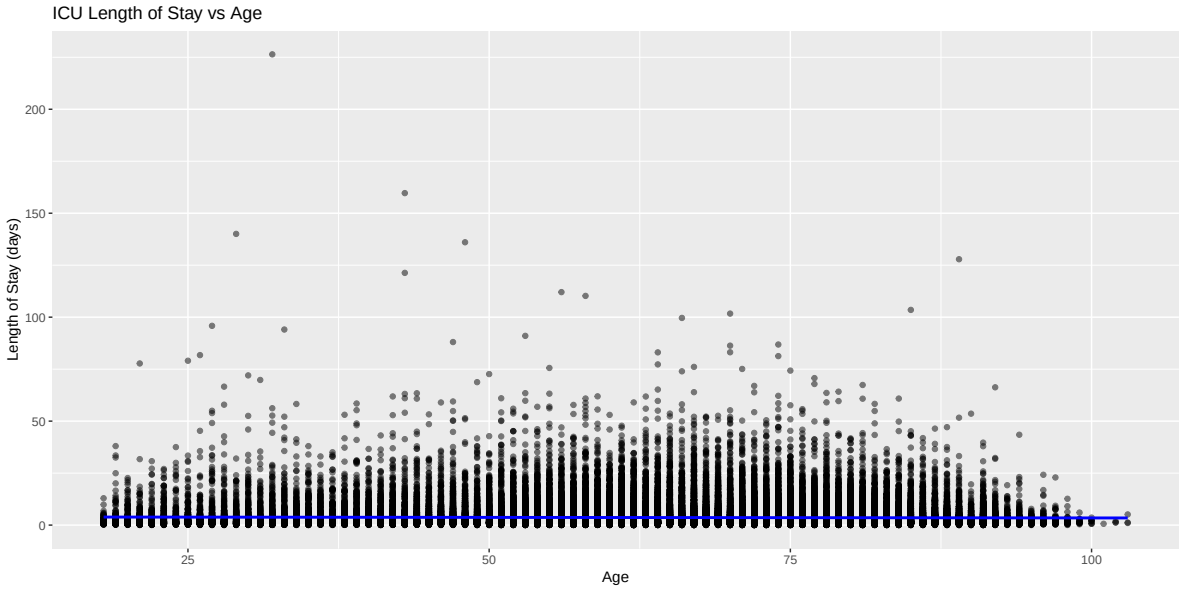



```
ggplot(mimic_icu_cohort, aes(x = age_at_intime, y = los)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", color = "blue") +
```

```
  labs(title = "ICU Length of Stay vs Age", x = "Age",
       y = "Length of Stay (days)")
```

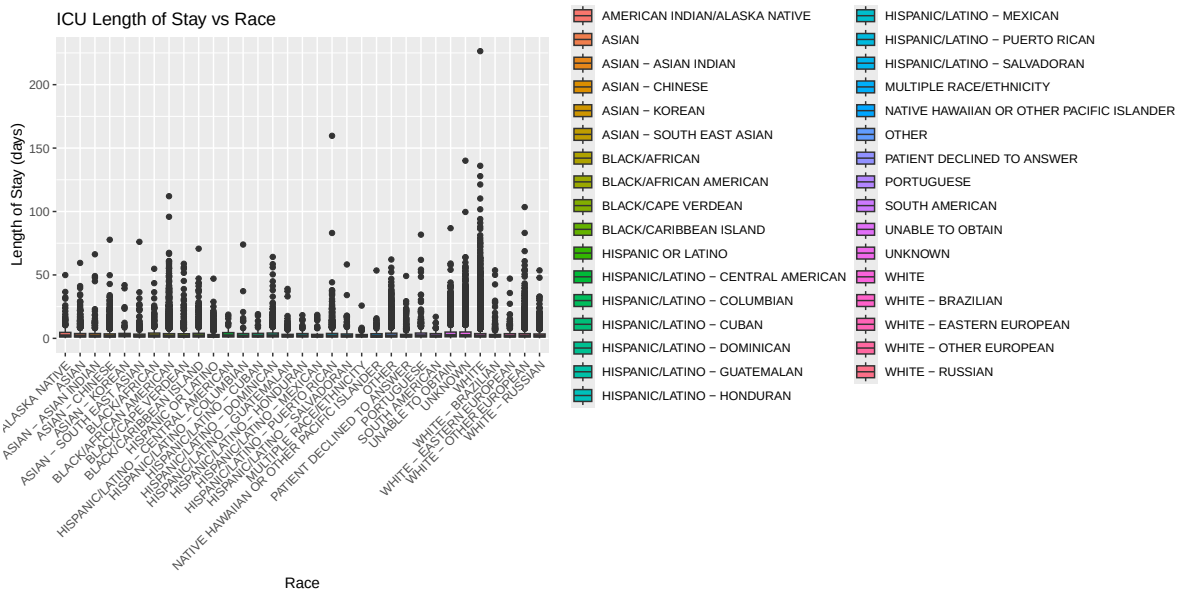
```
`geom_smooth()` using formula = 'y ~ x'
```

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



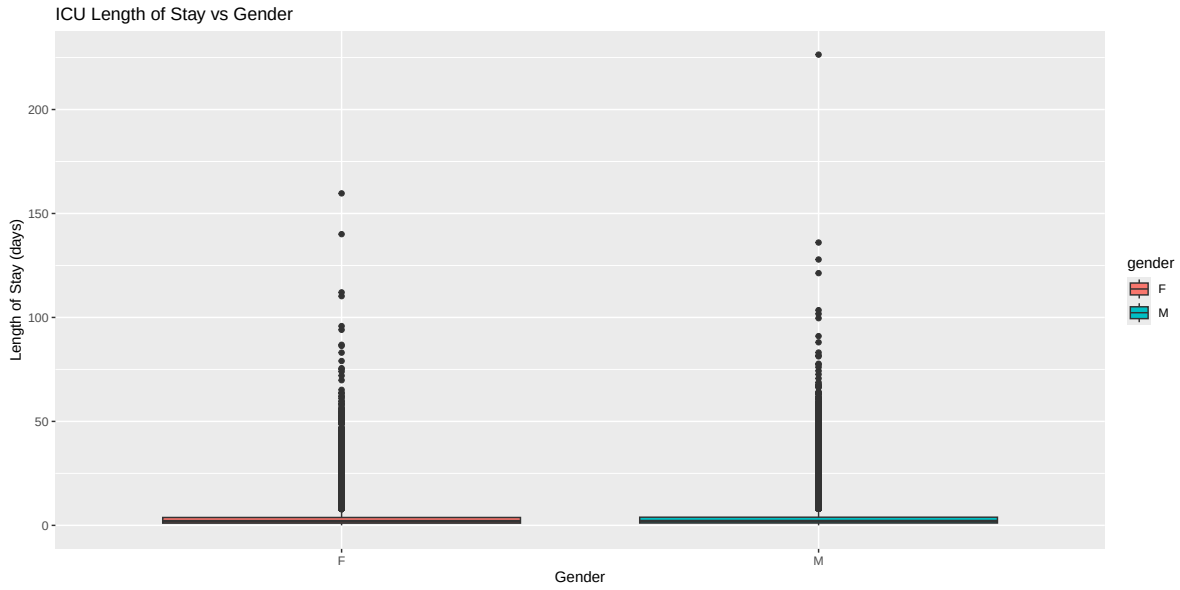

```
ggplot(mimic_icu_cohort, aes(x = race, y = los, fill = race)) +
  geom_boxplot() +
  labs(title = "ICU Length of Stay vs Race", x = "Race",
       y = "Length of Stay (days)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

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



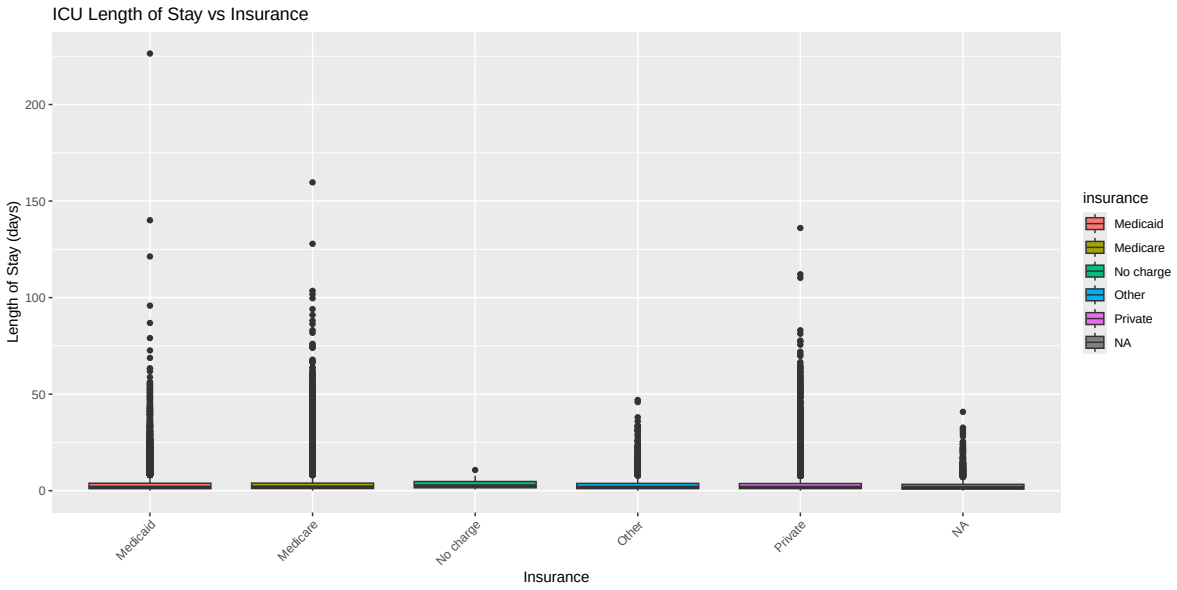

```
ggplot(mimic_icu_cohort, aes(x = gender, y = los, fill = gender)) +
  geom_boxplot() +
  labs(title = "ICU Length of Stay vs Gender", x = "Gender",
       y = "Length of Stay (days)")
```

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



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

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



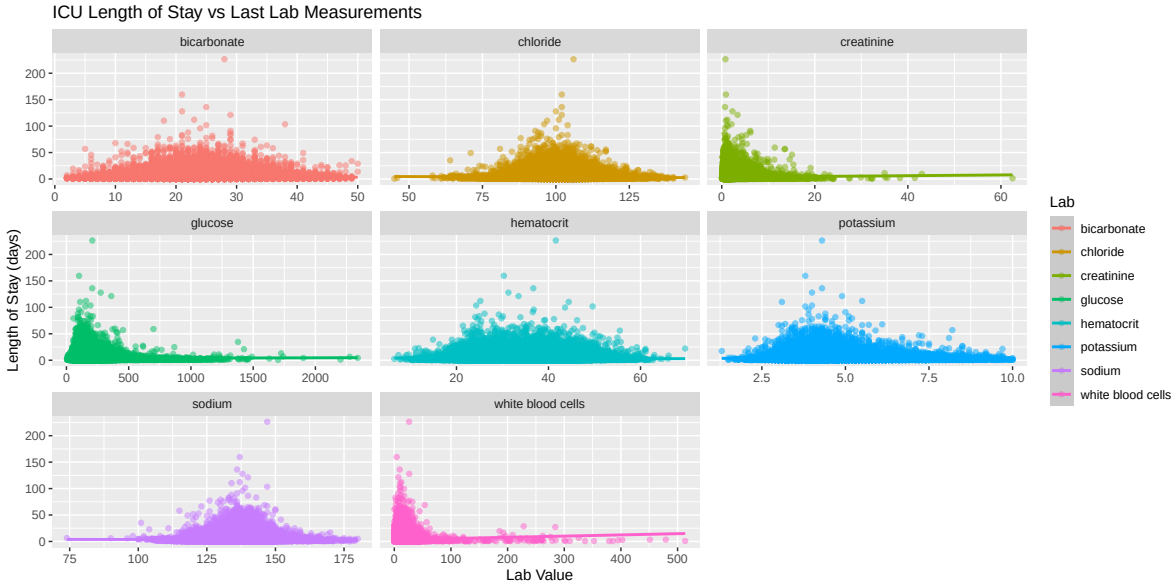
- Length of ICU stay `los` vs the last available lab measurements before ICU stay

```
mimic_icu_cohort %>%
  select(los, `white blood cells`, hematocrit, creatinine, sodium, glucose,
         potassium, chloride, bicarbonate) %>%
  pivot_longer(cols = -los, names_to = "Lab", values_to = "Value") %>%
  ggplot(aes(x = Value, y = los, color = Lab)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  facet_wrap(~ Lab, scales = "free_x") +
  labs(title = "ICU Length of Stay vs Last Lab Measurements",
       x = "Lab Value", y = "Length of Stay (days)")
```

```
`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

```
mimic_icu_cohort %>%
  select(los, heart_rate, non_invasive_blood_pressure_systolic,
         non_invasive_blood_pressure_diastolic,
         temperature_fahrenheit, respiratory_rate) %>%
  cor(use = "complete.obs")
```

```
                                            los  heart_rate
los                                1.0000000000 0.067042611
heart_rate                         0.0670426114 1.000000000
non_invasive_blood_pressure_systolic -0.0037026394 0.001906456
non_invasive_blood_pressure_diastolic  0.0005899571 0.015957059
temperature_fahrenheit              0.0062384099 0.061290942
respiratory_rate                    0.0886158884 0.353521841
                                   non_invasive_blood_pressure_systolic
los                                                        -0.003702639
heart_rate                                                  0.001906456
```

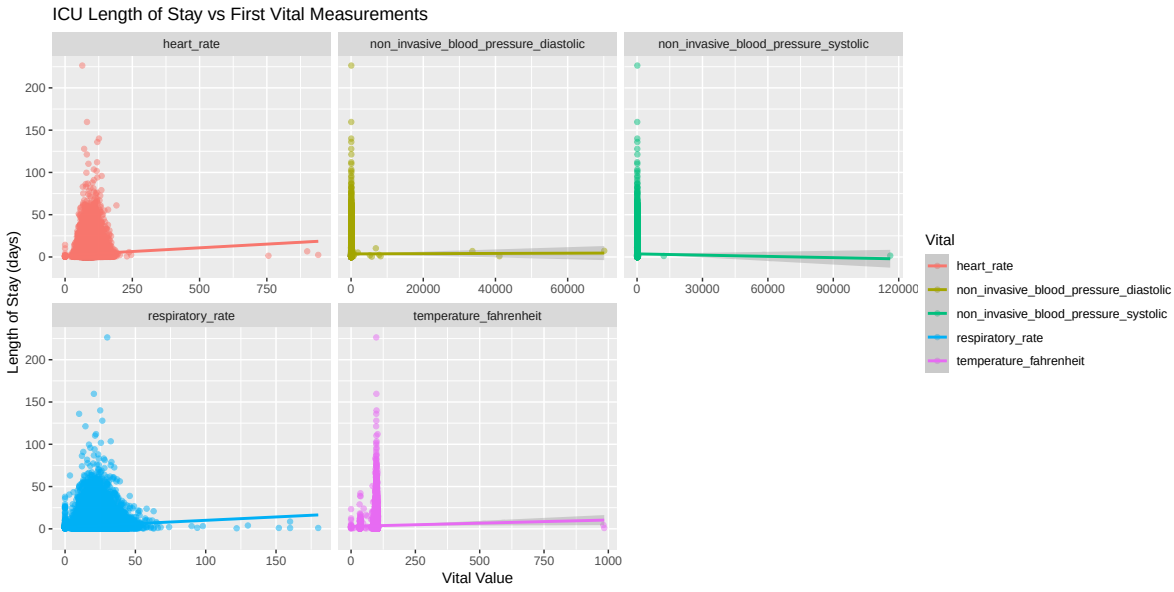
```
non_invasive_blood_pressure_systolic                          1.000000000
non_invasive_blood_pressure_diastolic                         0.002012290
temperature_fahrenheit                                        0.001348258
respiratory_rate                                              0.006628126
                                      non_invasive_blood_pressure_diastolic
los                                                           0.0005899571
heart_rate                                                    0.0159570593
non_invasive_blood_pressure_systolic                          0.0020122897
non_invasive_blood_pressure_diastolic                         1.0000000000
temperature_fahrenheit                                        0.0022057585
respiratory_rate                                              0.0119027496
                                      temperature_fahrenheit respiratory_rate
los                                               0.006238410      0.088615888
heart_rate                                        0.061290942      0.353521841
non_invasive_blood_pressure_systolic              0.001348258      0.006628126
non_invasive_blood_pressure_diastolic             0.002205759      0.011902750
temperature_fahrenheit                            1.000000000      0.037346642
respiratory_rate                                  0.037346642      1.000000000
```

```
mimic_icu_cohort %>%
  select(los, heart_rate, non_invasive_blood_pressure_systolic,
         non_invasive_blood_pressure_diastolic,
         temperature_fahrenheit, respiratory_rate) %>%
  pivot_longer(-los, names_to = "Vital", values_to = "Value") %>%
  ggplot(aes(x = Value, y = los, color = Vital)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm") +
  facet_wrap(~Vital, scales = "free_x") +
  labs(title = "ICU Length of Stay vs First Vital Measurements",
       x = "Vital Value", y = "Length of Stay (days)")
```

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

```
Warning: Removed 4774 rows containing non-finite outside the scale range
(`stat_smooth()`).
```

```
Warning: Removed 4774 rows containing missing values or values outside the scale range
(`geom_point()`).
```



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

```
mimic_icu_cohort %>%
  group_by(first_careunit) %>%
  summarise(mean_los = mean(los, na.rm = TRUE),
            median_los = median(los, na.rm = TRUE),
            sd_los = sd(los, na.rm = TRUE),
            count = n())
```

```
# A tibble: 17 x 5
   first_careunit                                mean_los median_los sd_los count
   <chr>                                            <dbl>      <dbl>  <dbl> <int>
 1 Cardiac Vascular Intensive Care Unit (CVICU)      3.32       1.99   4.99 14771
 2 Coronary Care Unit (CCU)                         3.09       2.01   3.57 10775
 3 Intensive Care Unit (ICU)                        8.79       5.76  10.3     33
 4 Med/Surg                                         1.44       1.44  NA         1
 5 Medical Intensive Care Unit (MICU)               3.76       1.91   5.91 20703
 6 Medical/Surgical Intensive Care Unit (MICU/~     3.09       1.79   4.59 15449
 7 Medicine                                        15.8       13.8   11.4     16
 8 Medicine/Cardiology Intermediate                 2.58       2.58  NA         1
 9 Neuro Intermediate                               5.02       3.00   6.05  5776
10 Neuro Stepdown                                   4.07       2.20   5.30  1421
11 Neuro Surgical Intensive Care Unit (Neuro S~     4.48       2.24   6.46  1751
12 Neurology                                       28.2       28.2   NA         1
13 PACU                                             4.02       2.00   5.65   122
```

```
14 Surgery/Trauma                          10.6       11.6    6.63    10
15 Surgery/Vascular/Intermediate           15.7       13.7   12.2    145
16 Surgical Intensive Care Unit (SICU)      3.90       1.98   6.15 13009
17 Trauma SICU (TSICU)                      3.64       1.88   5.42 10474
```

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

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