Biostat 203B Homework 3

Due Feb 23 @ 11:59PM

Yanzi Sun 106183069

Table of contents

Display machine information for reproducibility:

```r
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS Sonoma 14.7.4

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

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

time zone: America/Los_Angeles
tzcode source: internal

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

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

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

```r
library(arrow)
```

```


Attaching package: 'arrow'


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

    timestamp
```

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

```


Attaching package: 'pryr'


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

    where
```

```r
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
```

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2

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

```r
library(ggplot2)
```

Display your machine memory.

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

```
Totalram:   16.000 GiB
Freeram:   105.703 MiB
```

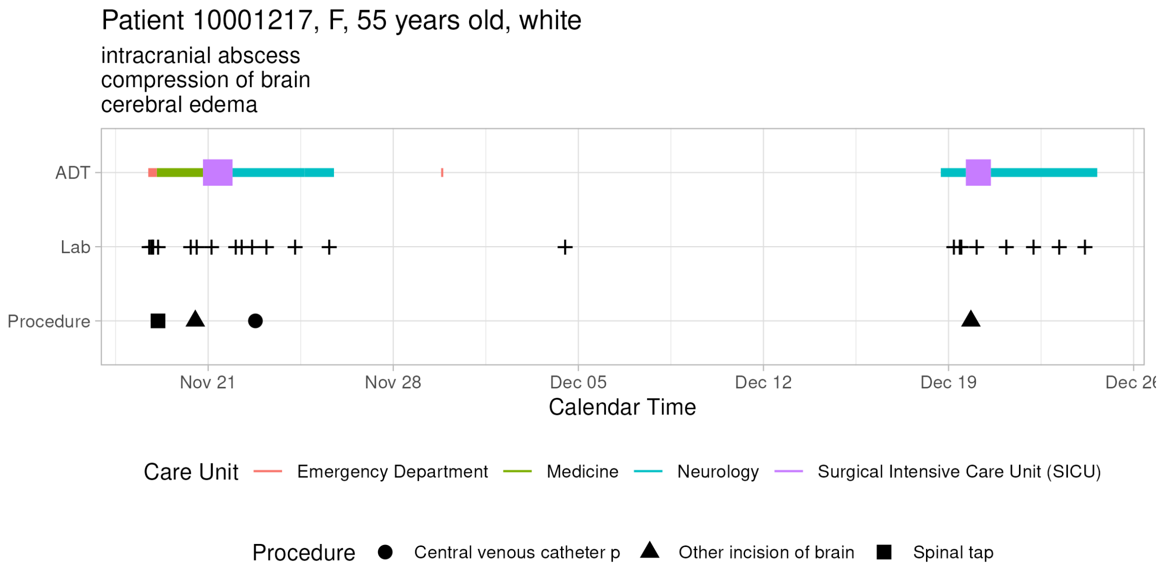
In this exercise, we use tidyverse (ggplot2, dplyr, etc) to explore the MIMIC-IV data introduced in homework 1 and to build a cohort of ICU stays.

Q1. Visualizing patient trajectory

Visualizing a patient's encounters in a health care system is a common task in clinical data analysis. In this question, we will visualize a patient's ADT (admission-discharge-transfer) history and ICU vitals in the MIMIC-IV data.

Q1.1 ADT history

A patient's ADT history records the time of admission, discharge, and transfer in the hospital. This figure shows the ADT history of the patient with `subject_id` 10001217 in the MIMIC-IV data. The x-axis is the calendar time, and the y-axis is the type of event (ADT, lab, procedure). The color of the line segment represents the care unit. The size of the line segment represents whether the care unit is an ICU/CCU. The crosses represent lab events, and the shape of the dots represents the type of procedure. The title of the figure shows the patient's demographic information and the subtitle shows top 3 diagnoses.



Do a similar visualization for the patient with `subject_id` 10063848 using ggplot.

Hint: We need to pull information from data files `patients.csv.gz`, `admissions.csv.gz`, `transfers.csv.gz`, `labevents.csv.gz`, `procedures_icd.csv.gz`, `diagnoses_icd.csv.gz`, `d_icd_procedures.csv.gz`, and `d_icd_diagnoses.csv.gz`. For the big file `labevents.csv.gz`, use the Parquet format you generated in Homework 2. For reproducibility, make the Parquet folder `labevents_pq` available at the current working directory `hw3`, for example, by a symbolic link. Make your code reproducible.

Solution: My result is shown below.

```
#use semi-join to filter & merge rows
labevents_pq <- arrow::open_dataset("labevents_parquet",
                                      format = "parquet") %>%
              filter(subject_id == 10063848) %>%
              collect()
```

```
#pulled data from&creating subjects patients_df, `admissions.csv.gz`, `transfers.csv.gz`, `la

patients_df <- read_csv("~/mimic/hosp/patients.csv.gz") |>
              filter(subject_id == 10063848) %>%
              collect()
```

```
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_df <- read_csv("~/mimic/hosp/admissions.csv.gz") |>
              filter(subject_id == 10063848) %>%
              collect()
```

```
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_df <- read_csv("~/mimic/hosp/transfers.csv.gz") |>
              filter(subject_id == 10063848) %>%
              collect()
```

```
Rows: 2413581 Columns: 7
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): eventtype, careunit
dbl  (3): subject_id, hadm_id, transfer_id
dttm (2): intime, outtime
```

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

```
procedures_icd_df <- read_csv("~/mimic/hosp/procedures_icd.csv.gz") |>
              filter(subject_id == 10063848) %>%
              collect()
```

```
Rows: 859655 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (1): icd_code
dbl  (4): subject_id, hadm_id, seq_num, icd_version
date (1): chartdate

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

```
diagnoses_icd_df <- read_csv("~/mimic/hosp/diagnoses_icd.csv.gz") |>
              filter(subject_id == 10063848) %>%
              collect()
```

```
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_df <- 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.
```

```r
d_icd_diagnoses_df <- 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.
```

```r
# Load necessary libraries
library(tidyverse)
library(lubridate)

# Convert date columns to Date format
transfers_data <- transfers_df |>
  mutate(intime = as.Date(intime), outtime = as.Date(outtime))

lab_data <- labevents_pq |>
  mutate(charttime = as.Date(charttime))

procedure_data <- procedures_icd_df |>
  mutate(chartdate = as.Date(chartdate))

diagnoses_translated <- diagnoses_icd_df |>
  left_join(d_icd_diagnoses_df, by = "icd_code") |>
  filter(icd_version.y == "10")  |>
  top_n(3, wt = seq_num)  # Select top 3 diagnoses
```

```
Warning in left_join(diagnoses_icd_df, d_icd_diagnoses_df, by = "icd_code"): Detected an une
i Row 17 of `x` matches multiple rows in `y`.
i Row 15793 of `y` matches multiple rows in `x`.
i If a many-to-many relationship is expected, set `relationship =
  "many-to-many"` to silence this warning.
```

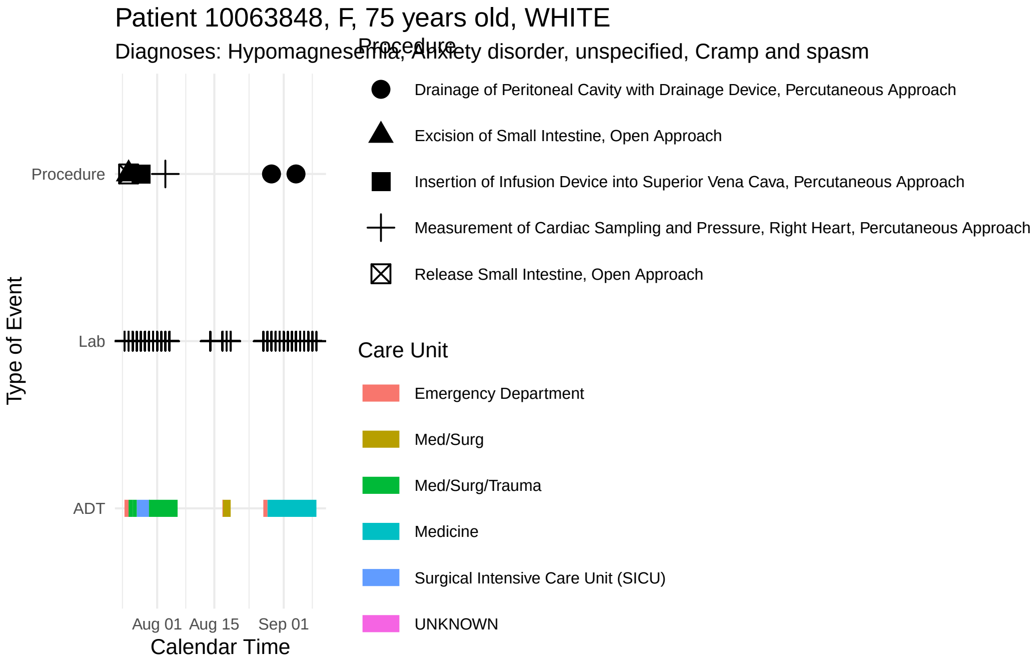
```r
# Convert to a readable format
diagnoses_text <- paste(diagnoses_translated$long_title,
                        collapse = ", ")
```

```r
# Merge procedures with descriptions
procedures_translated <- procedure_data  |>
  left_join(d_icd_procedures_df, by = "icd_code")

# Ensure procedure titles are available
procedure_data <- procedure_data  |>
  left_join(d_icd_procedures_df, by = "icd_code")

# Create ggplot visualization
ggplot() +
  # Plot Transfers (Care Units) as segments
  geom_segment(data = transfers_data,
               aes(x = intime, xend = outtime,
                   y = "ADT", yend = "ADT", color = careunit),
               size = 3) +

  # Plot Lab Events as points
  geom_point(data = lab_data,
             aes(x = charttime, y = "Lab"),
             shape = 3, size = 2) +

  # Plot Procedures as points with different shapes
  geom_point(data = procedure_data,
             aes(x = chartdate, y = "Procedure", shape = long_title),
             size = 3, fill = "black") +

  # Formatting
  labs(
    title = paste0("Patient ", patients_df$subject_id,
                   ", ", patients_df$gender, ", ",
                   patients_df$anchor_age, " years old, ",
                   admissions_df$race),
    subtitle = paste("Diagnoses:",
                     paste(diagnoses_text, collapse="\n")),
    x = "Calendar Time", y = "Type of Event",
    color = "Care Unit",
    shape = "Procedure"
  ) +
  scale_x_date(date_labels = "%b %d") +
  theme_minimal(base_size=8)
```

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

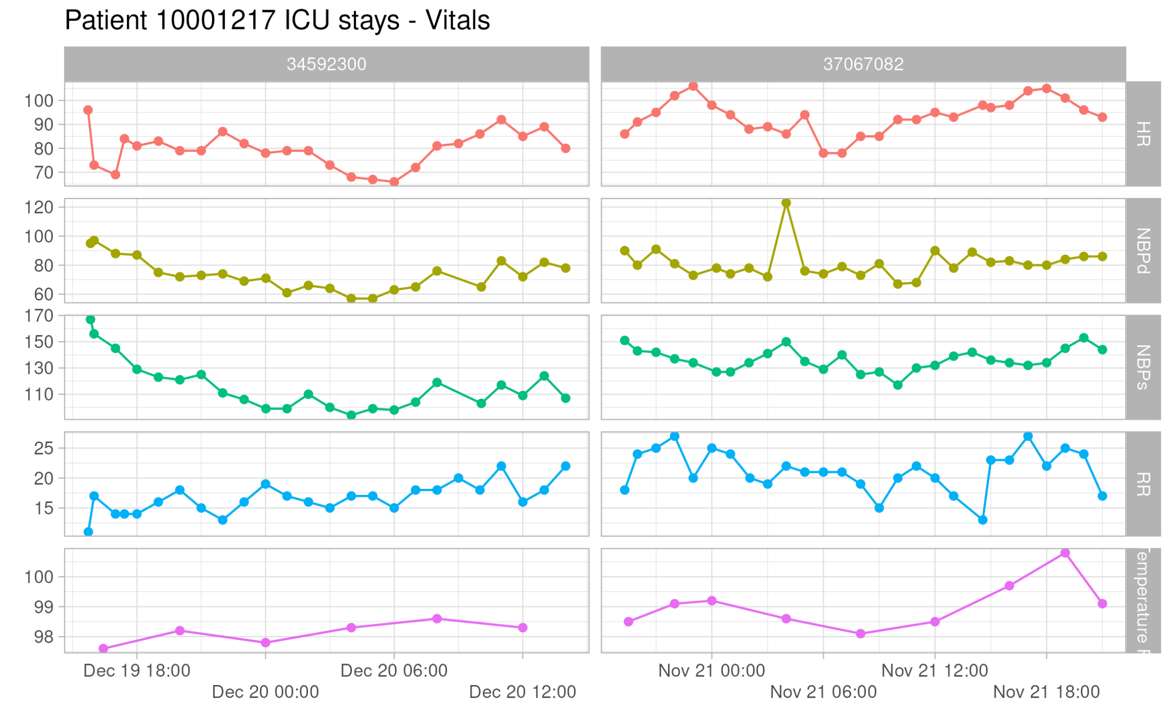
```
i Please use `linewidth` instead.
```

```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_segment()`).
```



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.

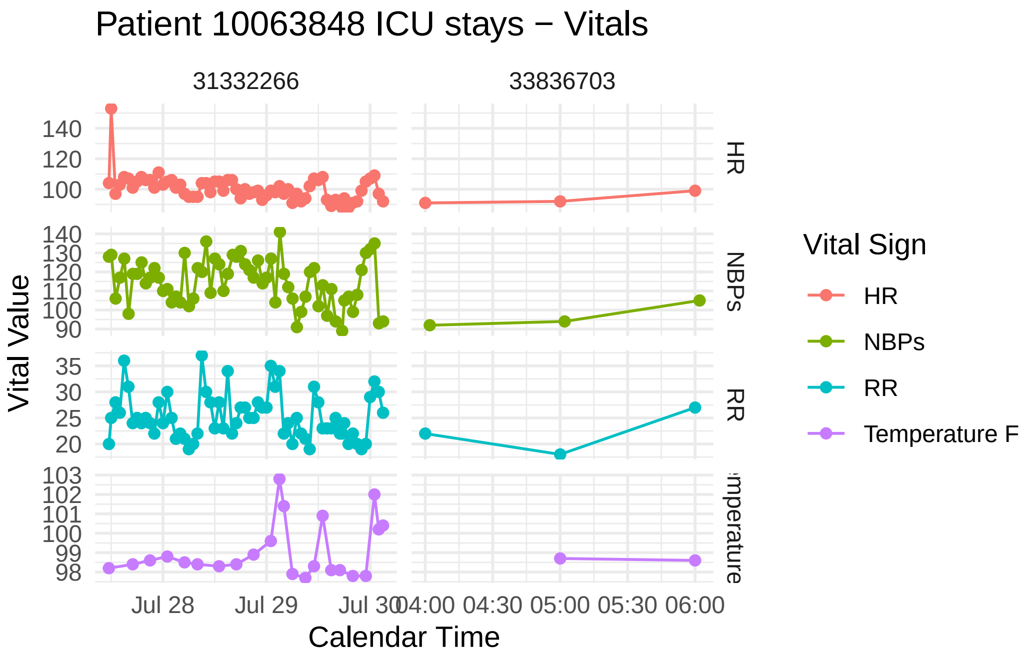
Solution: My work is shown below.

```r
#ingest all icu data
chartevents_pq <- arrow::open_dataset("chartevents_parquet",
                                      format = "parquet") |>
              filter(subject_id == 10063848,
                     itemid %in% c(220045, 220179, 223761, 220210)) |>
              collect()
d_items_df <- read_csv("~/mimic/icu/d_items.csv.gz") |>
            select(itemid, abbreviation)
```

```
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.
```

```
icu_all <- chartevents_pq |>
          left_join(y=d_items_df, by="itemid") |>
          select(subject_id, stay_id, charttime,
                 itemid, valuenum, abbreviation)
```

```
#I consulted with Bowen on this question and he said my computer may read time differently s
ggplot(icu_all, aes(x = charttime, y = valuenum, color = abbreviation )) +
  geom_line() +
  geom_point() +
  facet_grid(abbreviation ~ stay_id, scales = "free") +
  scale_x_datetime() +

  # Formatting
  labs(
    title = paste("Patient", "10063848", "ICU stays - Vitals"),
    x = "Calendar Time",
    y = "Vital Value",
    color = "Vital Sign"
  ) +

  # Improve theme aesthetics
  theme_minimal()
```



```
graphics.off()  # Closes all open plots
rm(list = ls()) # Clears all objects in the environment
gc()            # Runs garbage collection
```

```
          used (Mb) gc trigger  (Mb) limit (Mb) max used  (Mb)
Ncells 1823074 97.4    4172954 222.9         NA  4172954 222.9
Vcells 3417821 26.1   30402004 232.0      16384 47492245 362.4
```

Q2. ICU stays

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

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

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

Q2.1 Ingestion

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

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

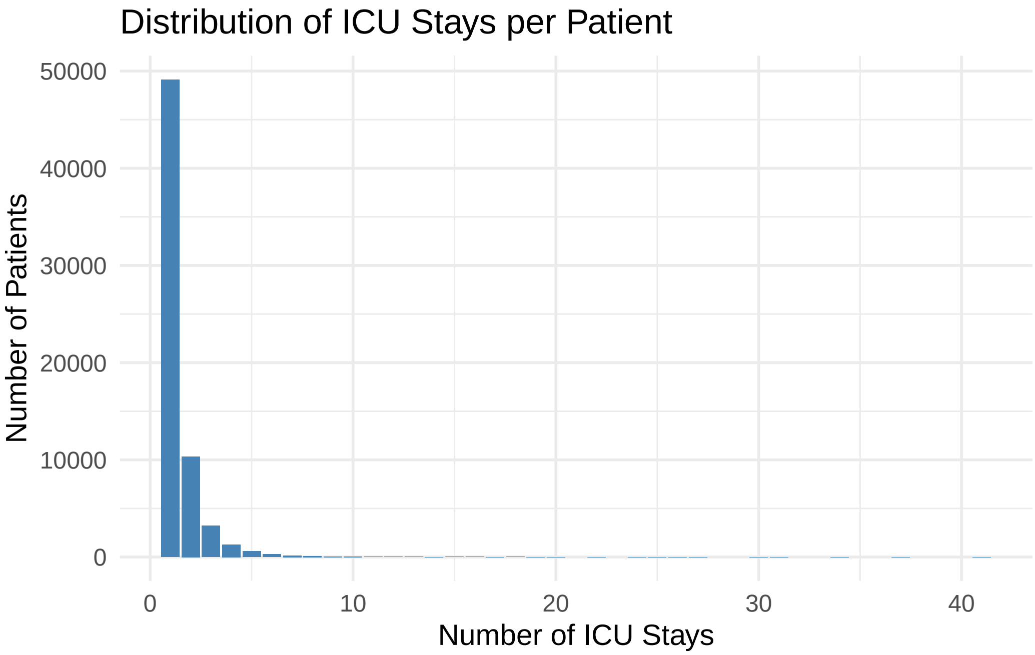
```
Rows: 94458 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): first_careunit, last_careunit
dbl  (4): subject_id, hadm_id, stay_id, los
dttm (2): intime, outtime

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.
```

Q2.2 Summary and visualization

How many unique `subject_id`? Can a `subject_id` have multiple ICU stays? Summarize the number of ICU stays per `subject_id` by graphs.

Solution: There are 65366 unique subject_id. Yes, a subject_id can have multiple icu stays as question 1 shows. The graph of #icu stays per subject_id is shown below.

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

stay_distribution <- icu_stay_counts %>%
  group_by(num_stays) %>%
  summarise(frequency = n())

# Create bar plot
ggplot(stay_distribution, aes(x = num_stays, y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(
    title = "Distribution of ICU Stays per Patient",
    x = "Number of ICU Stays",
    y = "Number of Patients"
  ) +
  theme_minimal()
```



Q3. `admissions` data

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

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

```
subject_id,hadm_id,admittime,dischtime,deathtime,admission_type,admit_provider_id,admission_l
10000032,22595853,2180-05-06 22:23:00,2180-05-07 17:15:00,,URGENT,P49AFC,TRANSFER FROM HOSPIT
10000032,22841357,2180-06-26 18:27:00,2180-06-27 18:49:00,,EW EMER.,P784FA,EMERGENCY ROOM,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 REF
10000108,27250926,2163-09-27 23:17:00,2163-09-28 09:04:00,,EU OBSERVATION,P40JML,EMERGENCY RO
10000117,22927623,2181-11-15 02:05:00,2181-11-15 14:52:00,,EU OBSERVATION,P47EY8,EMERGENCY RO
```

Q3.1 Ingestion

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

```r
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.
```

Q3.2 Summary and visualization

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

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

According to the MIMIC-IV documentation,

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

Solution: The graphs are plotted below.

In the number of admission per patient, I observed that most patients have only one admission, but one outlier patient has 40 admissions surprisingly.

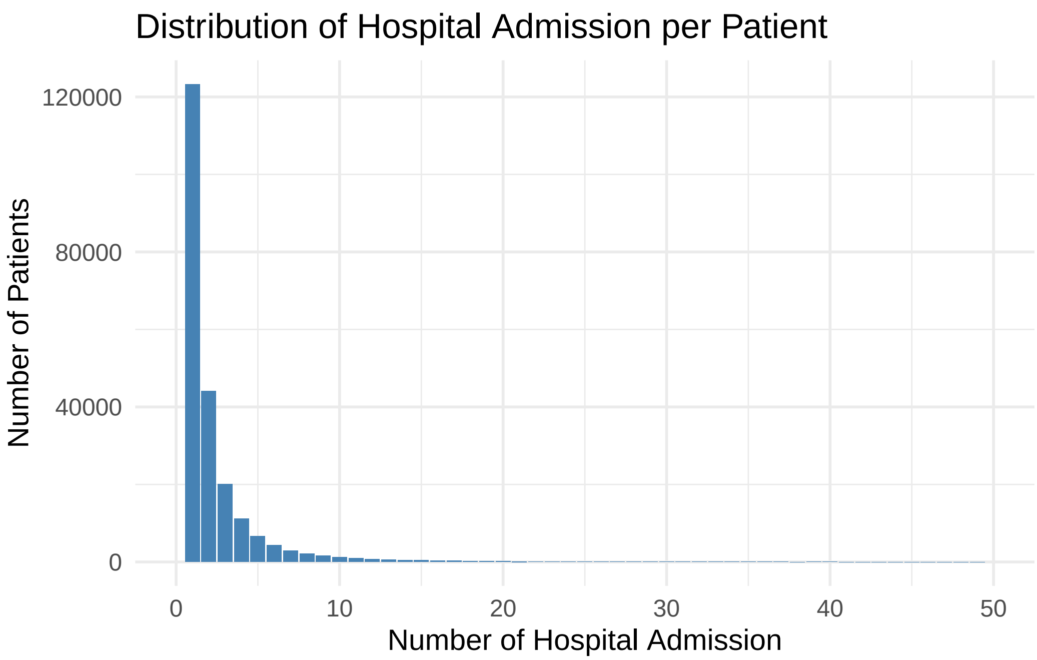
In the admission hour summary, the abnormality I found is the admission hour is mostly at 0am, which is the start of a day. I assume it is the default hour when logging admission time into the system.

Similarly in admission minute summary, I see most patients are admitted at 0, 15, 30, 45 minute, which can be due to the convienience to record the admission munite to the closest quarter point.

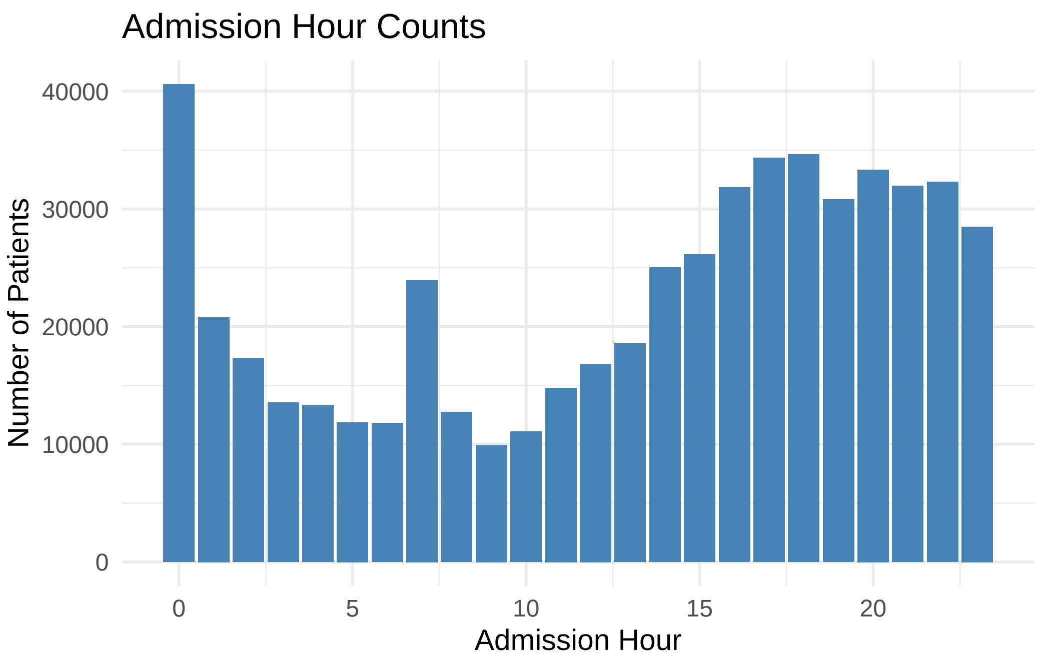
From the length of stay graph, most patients stay between 1-5 days. The abnormality I found is the distribution looks like a wave, with repeating peaks and troughs.The peaks are the integer hours and the troughs are hours with decimal. I think this is also due to the convienient-time-recording habit doctors have when inputting the time data.

```r
### number of admissions per patient
hosp_admission_counts <- admissions_tble %>%
  group_by(subject_id) %>%
  summarise(num_stays = n())

hosp_admission_distribution <- hosp_admission_counts %>%
  group_by(num_stays) %>%
  summarise(frequency = n())

ggplot(hosp_admission_distribution, aes(x = num_stays, y = frequency)) +
  xlim(0, 50)+
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(
    title = "Distribution of Hospital Admission per Patient",
    x = "Number of Hospital Admission",
    y = "Number of Patients"
  ) +
  theme_minimal()
```

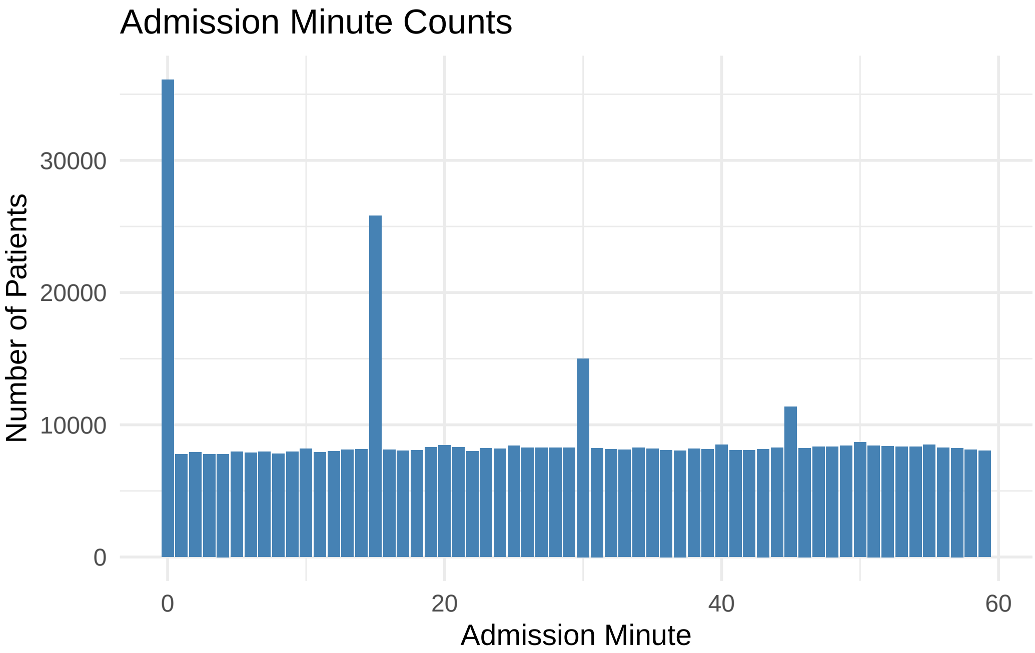
```
Warning: Removed 48 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



```
### admission hour (anything unusual?)
admission_hour <- admissions_tble %>%
  mutate(admit_hour = hour(admittime))

admission_hour_distribution <- admission_hour %>%
  group_by(admit_hour) %>%
  summarise(frequency = n())

ggplot(admission_hour_distribution, aes(x = admit_hour, y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
             labs(title = "Admission Hour Counts",
                  x = "Admission Hour",
                  y = "Number of Patients"
                  ) +
  theme_minimal()
```

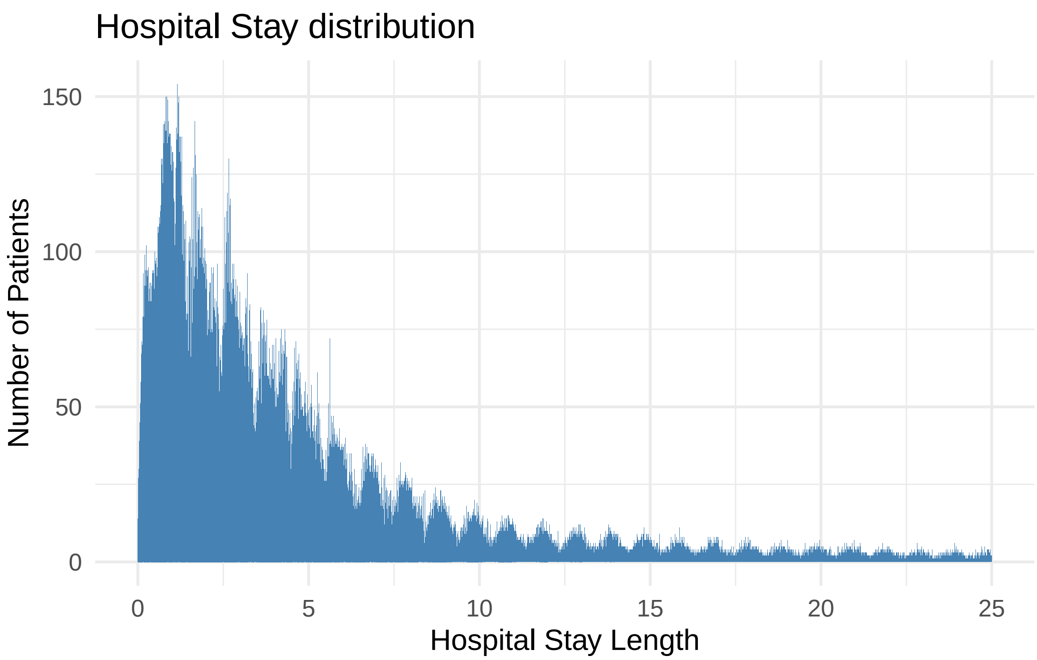


```
### admission minute (anything unusual?)
admission_minute <- admissions_tble %>%
  mutate(admit_minute = minute(admittime))

admission_minute_distribution <- admission_minute %>%
  group_by(admit_minute) %>%
  summarise(frequency = n())

ggplot(admission_minute_distribution, aes(x = admit_minute, y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
             labs(title = "Admission Minute Counts",
                  x = "Admission Minute",
                  y = "Number of Patients"
                  ) +
  theme_minimal()
```



```
### length of hospital stay (from admission to discharge) (anything unusual?)
hospital_stay <- admissions_tble |>
      mutate(length_of_stay = as.numeric(difftime(dischtime,
                                                  admittime,
                                                  units = "days")))

hosp_stay_distribution <- hospital_stay |>
  group_by(length_of_stay) |>
  summarise(frequency = n())

ggplot(hosp_stay_distribution, aes(x = length_of_stay, y = frequency)) +
  xlim(0, 25)+
  geom_bar(stat = "identity", fill = "steelblue") +
             labs(title = "Hospital Stay distribution",
                  x = "Hospital Stay Length",
                  y = "Number of Patients"
                  ) +
  theme_minimal()
```

```
Warning: Removed 9458 rows containing missing values or values outside the scale range
(`geom_bar()`).
```



Q4. `patients` data

Patient information is available in `patients.csv.gz`. See https://mimic.mit.edu/docs/iv/modules/hosp/patients/ for details of each field in this file. The first 10 lines are

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

```
subject_id,gender,anchor_age,anchor_year,anchor_year_group,dod
10000032,F,52,2180,2014 - 2016,2180-09-09
10000048,F,23,2126,2008 - 2010,
10000058,F,33,2168,2020 - 2022,
10000068,F,19,2160,2008 - 2010,
10000084,M,72,2160,2017 - 2019,2161-02-13
10000102,F,27,2136,2008 - 2010,
10000108,M,25,2163,2014 - 2016,
10000115,M,24,2154,2017 - 2019,
10000117,F,48,2174,2008 - 2010,
```

Q4.1 Ingestion

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

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

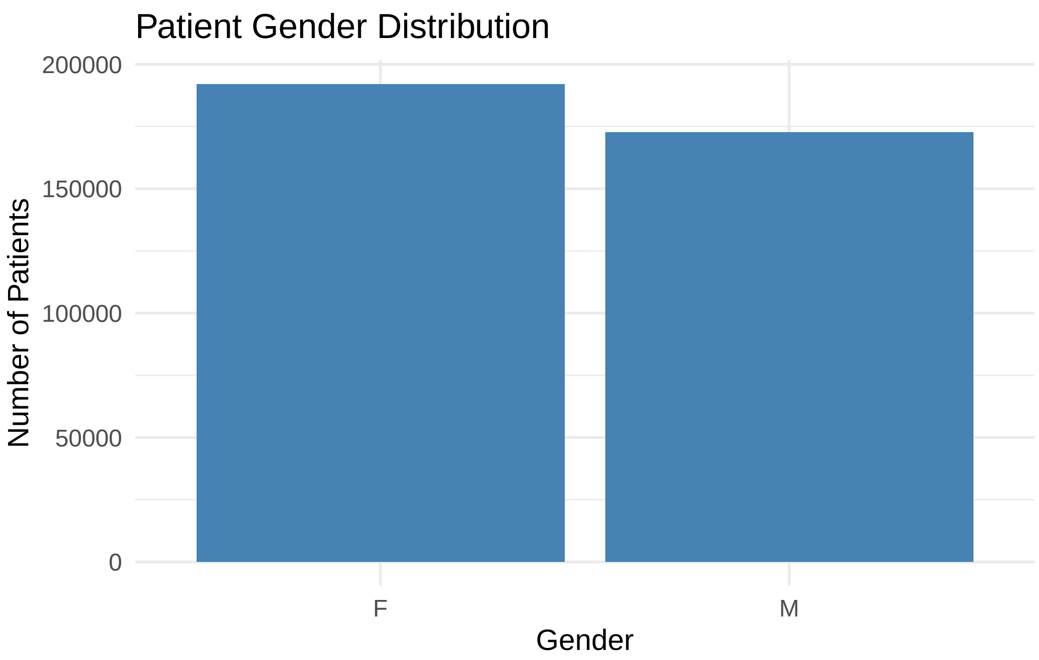
```
Rows: 364627 Columns: 6
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): gender, anchor_year_group
dbl  (3): subject_id, anchor_age, anchor_year
date (1): dod

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.
```

Q4.2 Summary and visualization

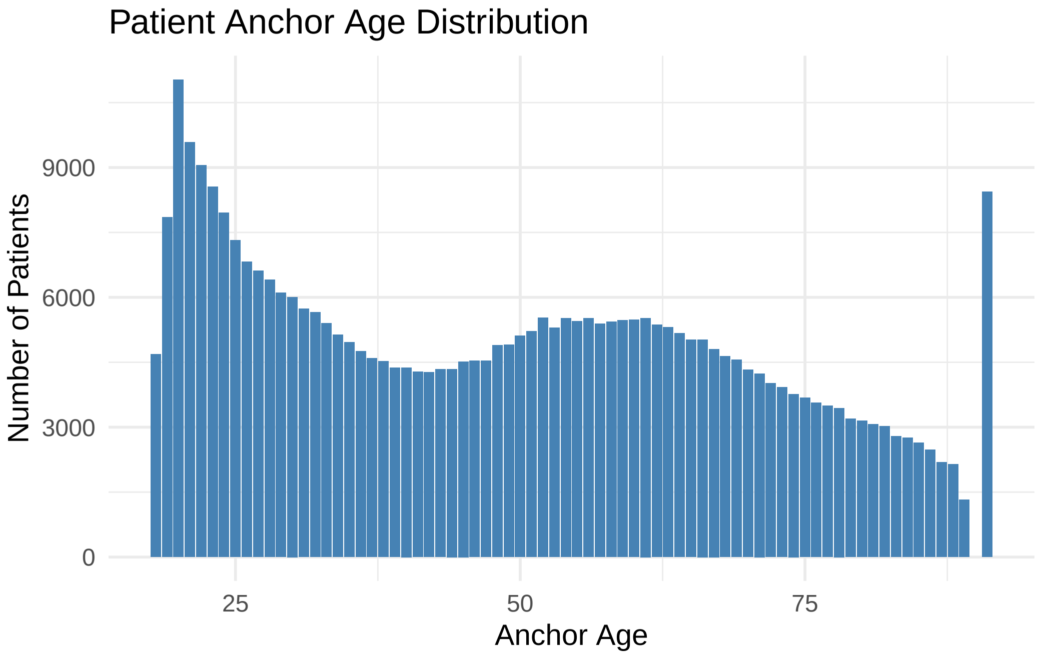
Summarize variables `gender` and `anchor_age` by graphics, and explain any patterns you see.

Solution: The summary graphics are drawn below. I see the gender distribution is fairly equal, with female slightly larger in number than male. The anchor age shows most patients are in their 20s and a lot of patients have the anchor age of ~90 yo, which is surprising but can be understandable due to the nature of the anchor age data protecting patients' privacy.

```r
### gender
gender_dist <- patients_tble |>
  group_by(gender) |>
  summarise(frequency = n())

age_dist <- patients_tble |>
  group_by(anchor_age) |>
  summarise(frequency = n())

ggplot(gender_dist, aes(x = gender, y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
             labs(title = "Patient Gender Distribution",
                  x = "Gender",
                  y = "Number of Patients"
                  ) +
  theme_minimal()
```



```r
ggplot(age_dist, aes(x = anchor_age, y = frequency)) +
  geom_bar(stat = "identity", fill = "steelblue") +
             labs(title = "Patient Anchor Age Distribution",
                  x = "Anchor Age",
                  y = "Number of Patients"
                  ) +
  theme_minimal()
```



Q5. Lab results

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

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

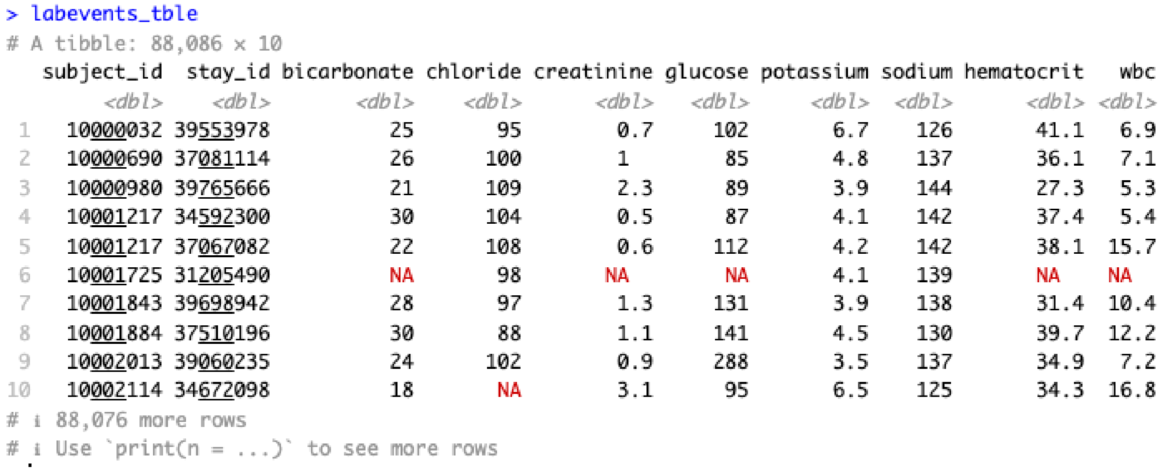
```
labevent_id,subject_id,hadm_id,specimen_id,itemid,order_provider_id,charttime,storetime,valu
1,10000032,,2704548,50931,P69FQC,2180-03-23 11:51:00,2180-03-23 15:56:00,___,95,mg/dL,70,100
2,10000032,,36092842,51071,P69FQC,2180-03-23 11:51:00,2180-03-23 16:00:00,NEG,,,,,,ROUTINE,
3,10000032,,36092842,51074,P69FQC,2180-03-23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



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

Hint: Use the Parquet format you generated in Homework 2. For reproducibility, make `labevents_pq` folder available at the current working directory `hw3`, for example, by a symbolic link. get the most recent lab result of each patients for each icu stay

Solution: My work is shown below.

```r
# I followed Dr.Zhou's instruction during lecture for this question
# create dictionary table
dlabitems_tble <- read.csv("~/mimic/hosp/d_labitems.csv.gz") |>
  filter(itemid %in% c(
    50912, 50971, 50983, 50902, 50882, 51221, 51301, 50931)) |>
  mutate(itemid = as.integer(itemid)) |>
  print()
```

```
  itemid             label fluid   category
1  50882       Bicarbonate Blood  Chemistry
2  50902          Chloride Blood  Chemistry
3  50912        Creatinine Blood  Chemistry
4  50931           Glucose Blood  Chemistry
5  50971         Potassium Blood  Chemistry
6  50983            Sodium Blood  Chemistry
7  51221        Hematocrit Blood Hematology
8  51301 White Blood Cells Blood Hematology
```

```r
# data wrangling step
labevents_tble <- open_dataset("labevents_parquet", format = "parquet") |>
  to_duckdb() |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in% dlabitems_tble$itemid) |>
  left_join(
    select(icustays_tble, subject_id, stay_id, intime),
    by=c("subject_id"),
    #copy=TRUE copies the r table into duckdb table to make them mergeable
    copy = TRUE )|>
  filter(storetime < intime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_max(storetime, n = 1) |>
  select(-storetime, -intime) |>
  ungroup() |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  rename_at (
    vars(as.character(dlabitems_tble$itemid)),
    ~str_to_lower(dlabitems_tble$label)
  ) |>
  rename(wbc = `white blood cells`) |>
  show_query() |>
  collect() |>
  arrange(subject_id, stay_id) |>
  relocate(subject_id, stay_id, chloride, hematocrit, bicarbonate, glucose, potassium, sodiu
  print(width = Inf)
```

```
<SQL>
SELECT
  subject_id,
  stay_id,
  MAX(CASE WHEN (itemid = 50912.0) THEN valuenum END) AS creatinine,
  MAX(CASE WHEN (itemid = 50983.0) THEN valuenum END) AS sodium,
  MAX(CASE WHEN (itemid = 50882.0) THEN valuenum END) AS bicarbonate,
  MAX(CASE WHEN (itemid = 50931.0) THEN valuenum END) AS glucose,
  MAX(CASE WHEN (itemid = 51221.0) THEN valuenum END) AS hematocrit,
  MAX(CASE WHEN (itemid = 50971.0) THEN valuenum END) AS potassium,
  MAX(CASE WHEN (itemid = 51301.0) THEN valuenum END) AS wbc,
  MAX(CASE WHEN (itemid = 50902.0) THEN valuenum END) AS chloride
FROM (
  SELECT subject_id, itemid, valuenum, stay_id
  FROM (
    SELECT
      q01.*,
      RANK() OVER (PARTITION BY subject_id, stay_id, itemid ORDER BY storetime DESC) AS col0
    FROM (
      SELECT LHS.*, stay_id, intime
      FROM (
        SELECT subject_id, itemid, storetime, valuenum
        FROM arrow_001
        WHERE (itemid IN (50882, 50902, 50912, 50931, 50971, 50983, 51221, 51301))
      ) LHS
      LEFT JOIN dbplyr_jn0OIuWHLx
        ON (LHS.subject_id = dbplyr_jn0OIuWHLx.subject_id)
    ) q01
    WHERE (storetime < intime)
  ) q01
  WHERE (col01 <= 1)
) q01
GROUP BY subject_id, stay_id
# A tibble: 88,086 x 10
   subject_id  stay_id chloride hematocrit bicarbonate glucose potassium sodium
        <dbl>    <dbl>    <dbl>      <dbl>       <dbl>   <dbl>     <dbl>  <dbl>
 1   10000032 39553978       95       41.1          25     102       6.7    126
 2   10000690 37081114      100       36.1          26      85       4.8    137
 3   10000980 39765666      109       27.3           21       89        3.9    144
 4   10001217 34592300      104       37.4           30       87        4.1    142
 5   10001217 37067082      108       38.1           22      112        4.2    142
 6   10001725 31205490       98       NA             NA       NA        4.1    139
 7   10001843 39698942       97       31.4           28      131        3.9    138
 8   10001884 37510196       88       39.7           30      141        4.5    130
 9   10002013 39060235      102       34.9           24      288        3.5    137
10   10002114 34672098       NA       34.3           18       95        6.5    125
     wbc creatinine
   <dbl>      <dbl>
 1   6.9        0.7
 2   7.1        1
 3   5.3        2.3
 4   5.4        0.5
 5  15.7        0.6
 6  NA         NA
 7  10.4        1.3
 8  12.2        1.1
 9   7.2        0.9
10  16.8        3.1
# i 88,076 more rows
```

Q6. Vitals from charted events

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

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

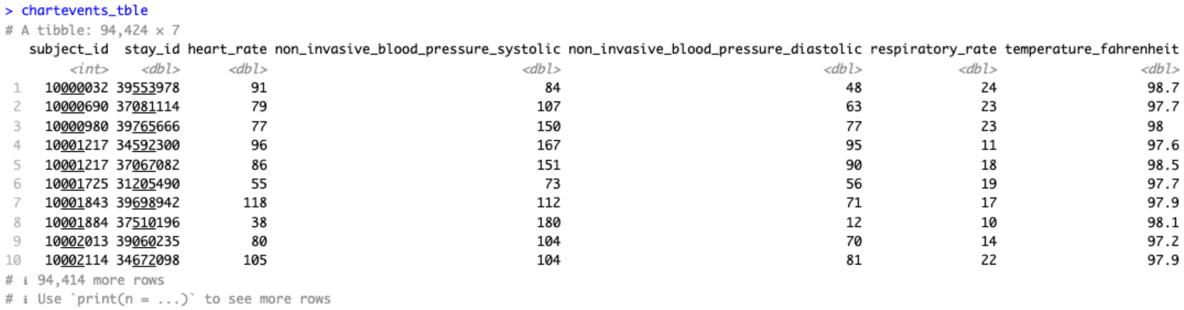
```
subject_id,hadm_id,stay_id,caregiver_id,charttime,storetime,itemid,value,valuenum,valueuom,wa
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226512,39.4,39.4,kg
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226707,60,60,Inch,0
10000032,29079034,39553978,18704,2180-07-23 12:36:00,2180-07-23 14:45:00,226730,152,152,cm,0
10000032,29079034,39553978,18704,2180-07-23 14:00:00,2180-07-23 14:18:00,220048,SR (Sinus Rh
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,highnormalvalu
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,6
220052,Arterial Blood Pressure mean,ABPm,chartevents,Routine Vital Signs,mmHg,Numeric,,
```

We are interested in the vitals for ICU patients: heart rate (220045), systolic non-invasive blood pressure (220179), diastolic non-invasive blood pressure (220180), body temperature in Fahrenheit (223761), and respiratory rate (220210). Retrieve a subset of chartevents.csv.gz only containing these items for the patients in icustays_tble. Further restrict to the first vital measurement within the ICU stay. The final chartevents_tble should have one row per ICU stay and columns for each vital measurement.



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

Solution: My work is shown below.

```
#take the average of the value at all storetime

dchartitems_tble <- read.csv("~/mimic/icu/d_items.csv.gz") |>
  filter(itemid %in% c(
    220045, 220179, 220180, 223761, 220210)) |>
  mutate(itemid = as.integer(itemid)) |>
  print()
  itemid                                 label  abbreviation     linksto
1 220045                            Heart Rate            HR chartevents
2 220179  Non Invasive Blood Pressure systolic          NBPs chartevents
3 220180 Non Invasive Blood Pressure diastolic          NBPd chartevents
4 220210                      Respiratory Rate            RR chartevents
5 223761                Temperature Fahrenheit Temperature F chartevents
             category unitname param_type lownormalvalue highnormalvalue
1 Routine Vital Signs      bpm    Numeric             NA              NA
2 Routine Vital Signs     mmHg    Numeric             NA              NA
3 Routine Vital Signs     mmHg    Numeric             NA              NA
4         Respiratory insp/min    Numeric             NA              NA
5 Routine Vital Signs       °F    Numeric             NA              NA
```

```
chartevents_tble <- open_dataset("chartevents_parquet", format = "parquet") |>
  to_duckdb() |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in% dchartitems_tble$itemid) |>
  left_join(
    select(icustays_tble, subject_id, stay_id),
    by=c("subject_id"),
    #copy=TRUE copies the r table into duckdb table to make them mergeable
    copy = TRUE )|>
  group_by(subject_id, stay_id, itemid) |>
  # i forgot if Dr. Zhou want us to take average of the mean value or use the first stored va
  #summarise(mean_valuenum = mean(valuenum, na.rm = TRUE), .groups = "drop") |>
  slice_min(storetime, n = 1) |>
  select(-storetime) |>
  ungroup() |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  rename_at (
    vars(as.character(dchartitems_tble$itemid)),
    ~str_to_lower(dchartitems_tble$label)
  ) |>
  # # # show_query() |>
  collect() |>
  arrange(subject_id, stay_id) |>
  relocate(subject_id, stay_id, `heart rate`,
           `non invasive blood pressure diastolic`,
           `non invasive blood pressure systolic`,
           `respiratory rate`, `temperature fahrenheit`) |>
  print(width = Inf)
```

```
# A tibble: 94,458 x 7
   subject_id  stay_id `heart rate` `non invasive blood pressure diastolic`
        <dbl>    <dbl>        <dbl>                                   <dbl>
 1   10000032 39553978           91                                      48
 2   10000690 37081114           80                                      63
 3   10000980 39765666           77                                     127
 4   10001217 34592300           86                                      90
 5   10001217 37067082           86                                      90
 6   10001725 31205490           86                                      56
 7   10001843 39698942          131                                      85
 8   10001884 37510196           60                                      49
 9   10002013 39060235           80                                      70
10   10002114 34672098          111                                      80
   `non invasive blood pressure systolic` `respiratory rate`
                                    <dbl>              <dbl>
 1                                     84                 24
 2                                    107                 27
 3                                    158                 24
 4                                    151                 18
 5                                    151                 18
 6                                     73                 19
 7                                    112                 17
 8                                    180                 16
 9                                    104                 14
10                                    112                 22
   `temperature fahrenheit`
                      <dbl>
 1                     98.7
 2                     97.7
 3                     98
 4                     98.5
 5                     98.5
 6                     97.7
 7                     97.9
 8                     98.1
 9                     97.2
10                     97.9
# i 94,448 more rows
```

Q7. Putting things together

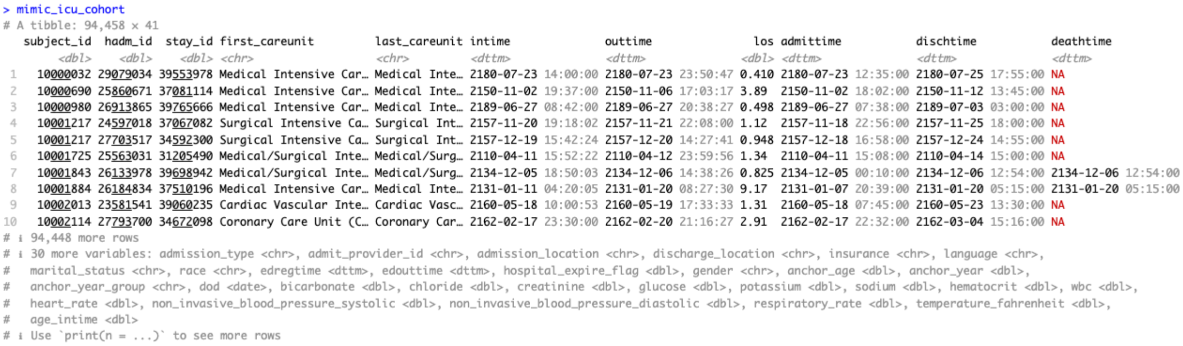
Let us create a tibble `mimic_icu_cohort` for all ICU stays, where rows are all ICU stays of adults (age at `intime` >= 18) and columns contain at least following variables

- all variables in `icustays_tble`

- all variables in `admissions_tble`

- all variables in `patients_tble`
- the last lab measurements before the ICU stay in `labevents_tble`
- the first vital measurements during the ICU stay in `chartevents_tble`

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



```
> mimic_icu_cohort
# A tibble: 94,458 × 41
   subject_id  hadm_id  stay_id first_careunit         last_careunit intime              outtime               los admittime           dischtime           deathtime
        <dbl>    <dbl>    <dbl> <chr>                  <chr>         <dttm>              <dttm>              <dbl> <dttm>              <dttm>              <dttm>
 1   10000032 29079034 39553978 Medical Intensive Car… Medical Inte… 2180-07-23 14:00:00 2180-07-23 23:50:47 0.410 2180-07-23 12:35:00 2180-07-25 17:55:00 NA
 2   10000690 25860671 37081114 Medical Intensive Car… Medical Inte… 2150-11-02 19:37:00 2150-11-06 17:03:17 3.89  2150-11-02 18:02:00 2150-11-12 13:45:00 NA
 3   10000980 26913865 39765666 Medical Intensive Car… Medical Inte… 2189-06-27 08:42:00 2189-06-27 20:38:27 0.498 2189-06-27 07:38:00 2189-07-03 03:00:00 NA
 4   10001217 24597018 37067082 Surgical Intensive Ca… Surgical Int… 2157-11-20 19:18:02 2157-11-21 22:08:00 1.12  2157-11-18 22:56:00 2157-11-25 18:00:00 NA
 5   10001217 27703517 34592300 Surgical Intensive Ca… Surgical Int… 2157-12-19 15:42:24 2157-12-20 14:27:41 0.948 2157-12-18 16:58:00 2157-12-24 14:55:00 NA
 6   10001725 25563031 31205490 Medical/Surgical Inte… Medical/Surg… 2110-04-11 15:52:22 2110-04-12 23:59:56 1.34  2110-04-11 15:08:00 2110-04-14 15:00:00 NA
 7   10001843 26133978 39698942 Medical/Surgical Inte… Medical/Surg… 2134-12-05 18:50:03 2134-12-06 14:38:26 0.825 2134-12-05 00:10:00 2134-12-06 12:54:00 2134-12-06 12:54:00
 8   10001884 26184834 37510196 Medical Intensive Car… Medical Inte… 2131-01-11 04:20:05 2131-01-20 08:27:30 9.17  2131-01-07 20:39:00 2131-01-20 05:15:00 2131-01-20 05:15:00
 9   10002013 23581541 39060235 Cardiac Vascular Inte… Cardiac Vasc… 2160-05-18 10:00:53 2160-05-19 17:33:33 1.31  2160-05-18 07:45:00 2160-05-23 13:30:00 NA
10   10002114 27793700 34672098 Coronary Care Unit (C… Coronary Car… 2162-02-17 23:30:00 2162-02-20 21:16:27 2.91  2162-02-17 22:32:00 2162-03-04 15:16:00 NA
# ℹ 94,448 more rows
# ℹ 30 more variables: admission_type <chr>, admit_provider_id <chr>, admission_location <chr>, discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dttm>, edouttime <dttm>, hospital_expire_flag <dbl>, gender <chr>, anchor_age <dbl>, anchor_year <dbl>,
#   anchor_year_group <chr>, dod <date>, bicarbonate <dbl>, chloride <dbl>, creatinine <dbl>, glucose <dbl>, potassium <dbl>, sodium <dbl>, hematocrit <dbl>, wbc <dbl>,
#   heart_rate <dbl>, non_invasive_blood_pressure_systolic <dbl>, non_invasive_blood_pressure_diastolic <dbl>, respiratory_rate <dbl>, temperature_fahrenheit <dbl>,
#   age_intime <dbl>
# ℹ Use `print(n = ...)` to see more rows
```

Solution: My work is shown below.

```r
#according to mimic documentation online, age of a patient = hospital admission time - ancho
mimic_icu_cohort <- patients_tble |>
  left_join(icustays_tble, by = c("subject_id")) |>
  mutate(age_at_intime = year(intime) - anchor_year + anchor_age) |>
  filter(age_at_intime >=18) |>
  left_join(admissions_tble, by = c("subject_id","hadm_id"))

first_vitals <- chartevents_tble |>
  group_by(subject_id, stay_id) |>
  slice_min(stay_id, n = 1) |>  # Get the first recorded value
  ungroup()

#  Summarize LAST lab values BEFORE ICU stay
last_labs <- labevents_tble |>
  group_by(subject_id, stay_id) |>
  slice_max(stay_id, n = 1) |>  # Get the last recorded value before ICU
  ungroup()

#  JOIN summarized tables to mimic_icu_cohort
mimic_icu_cohort <- mimic_icu_cohort |>
  left_join(first_vitals, by = c("subject_id", "stay_id")) |>
  left_join(last_labs, by = c("subject_id", "stay_id"))

# Check final structure
print(mimic_icu_cohort)
```

```
# A tibble: 94,458 x 41
   subject_id gender anchor_age anchor_year anchor_year_group dod        hadm_id
        <dbl> <chr>       <dbl>       <dbl> <chr>             <date>       <dbl>
 1   10000032 F              52        2180 2014 - 2016       2180-09-09  2.91e7
 2   10000690 F              86        2150 2008 - 2010       2152-01-30  2.59e7
 3   10000980 F              73        2186 2008 - 2010       2193-08-26  2.69e7
 4   10001217 F              55        2157 2011 - 2013       NA          2.46e7
 5   10001217 F              55        2157 2011 - 2013       NA          2.77e7
 6   10001725 F              46        2110 2011 - 2013       NA          2.56e7
 7   10001843 M              73        2131 2017 - 2019       2134-12-06  2.61e7
 8   10001884 F              68        2122 2008 - 2010       2131-01-20  2.62e7
 9   10002013 F              53        2156 2008 - 2010       NA          2.36e7
10   10002114 M              56        2162 2020 - 2022       2162-12-11  2.78e7
# i 94,448 more rows
# i 34 more variables: stay_id <dbl>, first_careunit <chr>,
#   last_careunit <chr>, intime <dttm>, outtime <dttm>, los <dbl>,
#   age_at_intime <dbl>, admittime <dttm>, dischtime <dttm>, deathtime <dttm>,
#   admission_type <chr>, admit_provider_id <chr>, admission_location <chr>,
#   discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dttm>, edouttime <dttm>, ...
```

Q8. Exploratory data analysis (EDA)

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

- Length of ICU stay `los` vs demographic variables (race, insurance, marital_status, gender, age at intime)
- Length of ICU stay `los` vs the last available lab measurements before ICU stay
- Length of ICU stay `los` vs the first vital measurements within the ICU stay

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

Solution: I choose to do a numeric summary for the first 3 and a ggplot for the last one. The results are shown below.

```r
demographics <- c("race", "insurance", "marital_status", "gender", "age_at_intime")
for (i in demographics) {
  summary_table <- mimic_icu_cohort |>
  group_by(.data[[i]]) |>
  summarise(mean_los = mean(los, na.rm = TRUE),
            median_los = median(los, na.rm = TRUE)) |>
  arrange(desc(mean_los)) |>
    print()
}
```

```
# A tibble: 33 x 3
   race                          mean_los median_los
   <chr>                            <dbl>      <dbl>
 1 UNABLE TO OBTAIN                  4.72       2.36
 2 UNKNOWN                           4.52       2.27
 3 ASIAN - KOREAN                    4.44       2.25
 4 PORTUGUESE                        4.41       2.14
 5 BLACK/CARIBBEAN ISLAND            4.34       2.04
 6 AMERICAN INDIAN/ALASKA NATIVE     4.31       2.08
 7 HISPANIC/LATINO - COLUMBIAN       4.10       1.80
 8 HISPANIC/LATINO - DOMINICAN       4.10       2.13
 9 ASIAN - ASIAN INDIAN              4.08       1.90
10 BLACK/AFRICAN                     4.01       2.08
# i 23 more rows
# A tibble: 6 x 3
  insurance mean_los median_los
  <chr>        <dbl>      <dbl>
1 No charge     3.87       2.60
2 Medicaid      3.79       1.90
3 Private       3.64       1.88
4 Medicare      3.60       2.03
5 Other         3.39       1.86
6 <NA>          3.21       1.65
# A tibble: 5 x 3
  marital_status mean_los median_los
  <chr>             <dbl>      <dbl>
1 <NA>               4.64       2.33
```

```
2 SINGLE             3.59       1.91
3 MARRIED            3.59       1.97
4 DIVORCED           3.58       1.95
5 WIDOWED            3.18       1.93
# A tibble: 2 x 3
  gender mean_los median_los
  <chr>     <dbl>      <dbl>
1 M          3.72       1.98
2 F          3.51       1.94
# A tibble: 86 x 3
   age_at_intime mean_los median_los
           <dbl>    <dbl>      <dbl>
 1            27     4.67       1.98
 2            58     4.09       2.02
 3            43     4.09       1.84
 4            32     4.08       1.83
 5            42     4.07       1.94
 6            47     4.02       2.04
 7            44     3.97       1.89
 8            67     3.97       2.03
 9            30     3.93       1.82
10            70     3.90       2.00
# i 76 more rows
```

```
lab_measurements <- names(labevents_tble)[-c(1,2)]
for (i in lab_measurements) {
  summary_table <- mimic_icu_cohort |>
  group_by(.data[[i]]) |>
  summarise(mean_los = mean(los, na.rm = TRUE),
            median_los = median(los, na.rm = TRUE)) |>
  arrange(desc(mean_los)) |>
    print()
}
```

```
# A tibble: 86 x 3
   chloride mean_los median_los
      <dbl>    <dbl>      <dbl>
 1       64     7.87       3.92
 2       66     6.29       4.77
 3       71     5.86       2.37
 4      124     5.47       3.21
 5      117     4.82       2.14
 6       NA     4.69       2.23
 7       88     4.60       2.20
 8      128     4.59       3.19
 9      123     4.43       2.71
10       73     4.38       3.72
# i 76 more rows
# A tibble: 526 x 3
   hematocrit mean_los median_los
        <dbl>    <dbl>      <dbl>
 1       69.7    22.1       22.1
 2       11      22.1       22.1
 3       61.1    18.9       18.9
 4       55.5    15.8        3.09
 5       11.2    12.4       12.4
 6        9.6    10.9       10.9
 7       53.1    10.6        6.24
 8       55      10.3        4.52
 9       63       9.85       9.85
10       55.7     9.81       1.93
# i 516 more rows
# A tibble: 58 x 3
   bicarbonate mean_los median_los
         <dbl>    <dbl>      <dbl>
 1        50      21.4       21.4
 2         8.6     8.80       8.80
 3        49       6.04       1.84
 4        41       5.94       3.11
 5        48       5.29       2.84
 6        45       5.11       2.63
 7        36       5.02       2.72
 8        38       4.96       2.65
 9        43       4.76       2.92
10        NA       4.68       2.21
# i 48 more rows
# A tibble: 954 x 3
   glucose mean_los median_los
     <dbl>    <dbl>      <dbl>
 1    1378     34.6      34.6
 2     454     29.1      29.1
 3    1426     20.9      20.9
 4     725     20.8      20.8
 5     699     20.8       2.98
 6     402     20.0      13.1
 7    1725     13.1      13.1
 8     858     11.9      11.9
 9    1225     10.9      10.9
10     360     10.9       2.01
# i 944 more rows
# A tibble: 89 x 3
   potassium mean_los median_los
       <dbl>    <dbl>      <dbl>
 1       1.3    17.5       17.5
 2       8.5     4.90       1.97
 3       9.7     4.82       1.98
 4      NA       4.69       2.23
 5       2.3     4.60       1.93
 6       8.1     4.53       2.21
 7       2.1     4.52       1.36
 8       9.2     4.35       2.75
 9       2.9     4.19       2.07
10       7.9     4.10       2.94
# i 79 more rows
# A tibble: 87 x 3
   sodium mean_los median_los
    <dbl>    <dbl>      <dbl>
 1    101    19.7       19.7
 2    104    10.6        5.12
 3     98    10.2       10.2
 4    102     8.65       8.65
 5     74     6.47       6.47
 6     90     6.06       6.06
 7     96     5.76       5.76
 8    159     5.65       2.87
 9    152     5.42       2.64
10    150     5.25       2.56
# i 77 more rows
# A tibble: 767 x 3
     wbc mean_los median_los
   <dbl>    <dbl>      <dbl>
 1  53.8     68.8      68.8
 2  44.9     57.0      57.0
 3  57.1     36.3      36.3
 4 228.      28.7      28.7
 5 284.      27.3      27.3
 6  53.6     21.2      21.2
 7 186.      19.9      19.9
 8  51.4     16.8      16.8
 9 193.      16.6      16.6
10  31.9     16.3       2.77
# i 757 more rows
# A tibble: 223 x 3
   creatinine mean_los median_los
        <dbl>    <dbl>      <dbl>
 1       13.6    20.3       11.8
 2       15.1    12.7        2.07
 3       35      11.3       11.3
 4       15.4    10.7        2.59
 5       43       9.74       9.74
 6       15.5     9.62       9.62
 7       19.1     8.47       2.61
 8        0.1     8.42       3.96
 9       14.6     7.60       3.23
10       23.2     7.22       7.22
# i 213 more rows
```

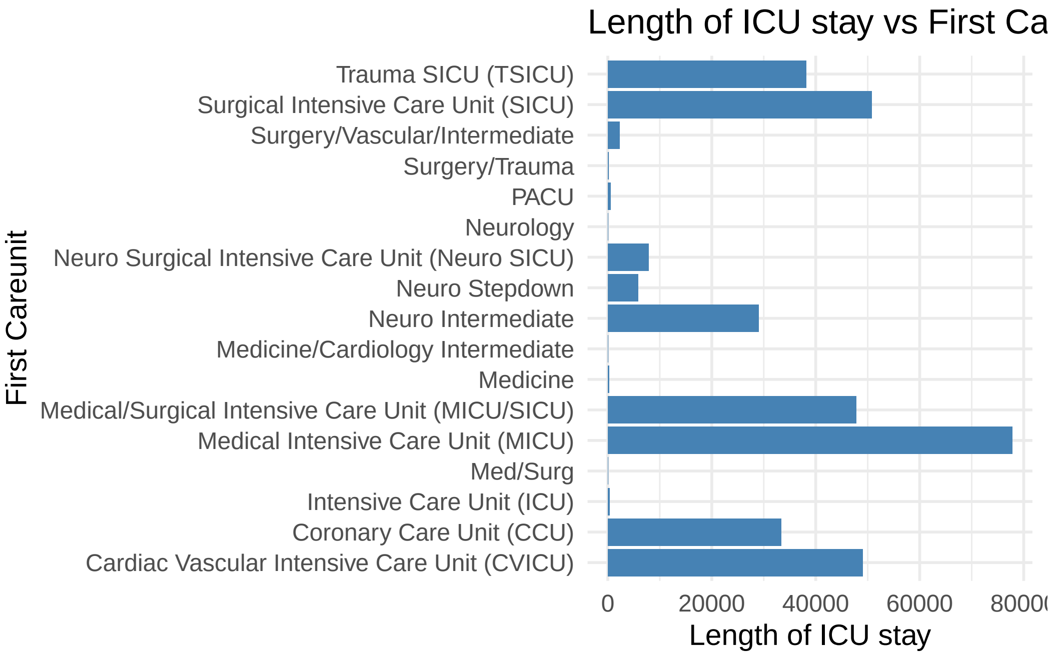
```r
vital_measurements <- names(chartevents_tble)[-c(1,2)]
for (i in vital_measurements) {
  summary_table <- mimic_icu_cohort |>
  group_by(.data[[i]]) |>
  summarise(mean_los = mean(los, na.rm = TRUE),
            median_los = median(los, na.rm = TRUE)) |>
  arrange(desc(mean_los)) |>
    print()
}
```

```
# A tibble: 173 x 3
   `heart rate` mean_los median_los
          <dbl>    <dbl>      <dbl>
 1          191    30.9       30.9
 2          167    10.7        6.75
 3          173     8.55       5.05
 4          180     6.27       2.58
 5          160     5.94       2.30
 6          179     5.94       2.91
 7          181     5.90       5.90
 8          171     5.75       1.68
 9          166     5.71       3.27
10          174     5.57       5.64
```

```
# i 163 more rows
# A tibble: 190 x 3
   `non invasive blood pressure diastolic` mean_los median_los
                                     <dbl>    <dbl>      <dbl>
 1                                     165    14.5       7.64
 2                                      18    11.2       1.74
 3                                     160    10.3       0.676
 4                                    6868     9.34      8.92
 5                                     153     7.94      2.67
 6                                     199     7.91      7.91
 7                                     174     7.81      4.97
 8                                   70130     7.33      7.33
 9                                    1052     6.53      6.53
10                                     156     6.37      5.78
# i 180 more rows
# A tibble: 214 x 3
   `non invasive blood pressure systolic` mean_los median_los
                                    <dbl>    <dbl>      <dbl>
 1                                    245    41.5       41.5
 2                                     56    16.1        6.49
 3                                     37    10.4       10.4
 4                                     75     8.22       3.44
 5                                    236     8.07       8.07
 6                                     58     7.97       4.74
 7                                     60     7.95       3.84
 8                                    240     7.01       7.01
 9                                     59     6.57       2.36
10                                     68     6.14       2.71
# i 204 more rows
# A tibble: 86 x 3
   `respiratory rate` mean_los median_los
                <dbl>    <dbl>      <dbl>
 1                 57     6.73       6.40
 2                 59     6.61       6.61
 3                 40     6.28       3.05
 4                115     6.20       6.20
 5                 63     6.12       1.63
 6                 58     5.92       5.92
 7                 85     5.64       5.64
 8                 95     5.36       5.36
 9                 75     5.32       5.32
10                 47     5.21       3.43
# i 76 more rows
# A tibble: 254 x 3
   `temperature fahrenheit` mean_los median_los
                      <dbl>    <dbl>      <dbl>
 1                     91.3     33.4      33.4 
 2                     85.5     31.7      31.7 
 3                     99.6     28.5      28.5 
 4                     82       18.7      18.7 
 5                     35.1     15.1       4.39
 6                     37.1     14.3       2.55
 7                     99.1     13.4      13.4 
 8                     90.9     12.9      12.9 
 9                     38.2     12.9      12.9 
10                    104.      12.1      12.1 
# i 244 more rows
```

```r
ggplot(mimic_icu_cohort, aes(x=first_careunit, y=los)) +
  geom_bar(stat = "identity", fill = "steelblue") +
   labs(
    title = "Length of ICU stay vs First Careunit",
    x = "First Careunit",
    y = "Length of ICU stay"
  ) +
  theme_minimal()+
  coord_flip()
```

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


Length of ICU stay vs First Ca
First Careunit
Trauma SICU (TSICU)
Surgical Intensive Care Unit (SICU)
Surgery/Vascular/Intermediate
Surgery/Trauma
PACU
Neurology
Neuro Surgical Intensive Care Unit (Neuro SICU)
Neuro Stepdown
Neuro Intermediate
Medicine/Cardiology Intermediate
Medicine
Medical/Surgical Intensive Care Unit (MICU/SICU)
Medical Intensive Care Unit (MICU)
Med/Surg
Intensive Care Unit (ICU)
Coronary Care Unit (CCU)
Cardiac Vascular Intensive Care Unit (CVICU)
0
20000
40000
60000
8000
Length of ICU stay