Biostat 203B Homework 3

Due Feb 21 @ 11:59PM

Khoa Vu 705600710

Display machine information for reproducibility:

```
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.1 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.0
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0

locale:
 [1] LC_CTYPE=C.UTF-8       LC_NUMERIC=C           LC_TIME=C.UTF-8
 [4] LC_COLLATE=C.UTF-8     LC_MONETARY=C.UTF-8    LC_MESSAGES=C.UTF-8
 [7] LC_PAPER=C.UTF-8       LC_NAME=C              LC_ADDRESS=C
[10] LC_TELEPHONE=C         LC_MEASUREMENT=C.UTF-8 LC_IDENTIFICATION=C

time zone: America/Los_Angeles
tzcode source: system (glibc)

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

loaded via a namespace (and not attached):
 [1] 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_1.0.3
```

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

```
library(arrow)
```

```
Attaching package: 'arrow'

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

    timestamp
```

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

```
Attaching package: 'pryr'

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

    where
```

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

```
Loading required package: R.oo

Loading required package: R.methodsS3

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

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

Attaching package: 'R.oo'

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

    throw
```

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

    getClasses, getMethods

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

    attach, detach, load, save

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


Attaching package: 'R.utils'

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

    timestamp

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

    timestamp

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

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

```
library(tidyverse)
```

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

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

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

Display your machine memory.

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

```
Totalram:  11.686 GiB
Freeram:   10.963 GiB
```

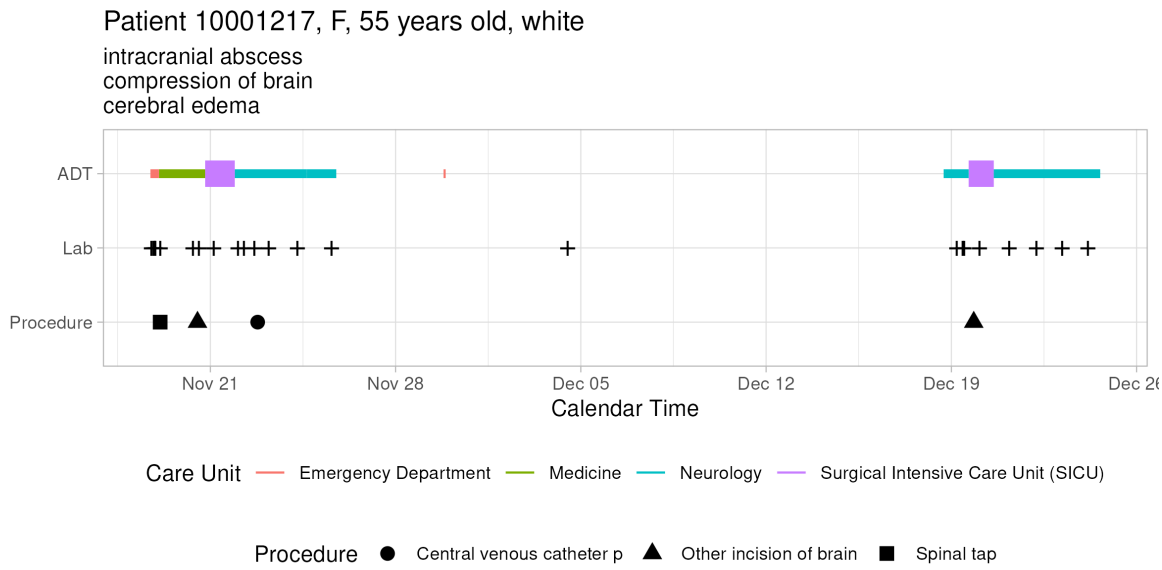
In this exercise, we use tidyverse (ggplot2, dplyr, etc) to explore the MIMIC-IV data introduced in homework 1 and to build a cohort of ICU stays.

Q1. Visualizing patient trajectory

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

Q1.1 ADT history

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



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

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

Solution:

```
#Setting the subject ID
#ID <- (10001217) #This is the exmaple subject ID
ID <- (10063848)

#Reading in labevents_pg, we only want the subject_id and charttime
labevents <- arrow::open_dataset("~/labevents_pq") %>%
  #Filering based on subject ID
  filter(subject_id %in% ID) %>%
  #Selecting only on columns of interest
  select(all_of(c("subject_id", "charttime"))) %>%
  #Renaming charttime to Calender_Time
  rename("Calender_Time" = charttime) %>%
  collect()
#Converting time to UTC and not PDT
labevents$Calender_Time <- as.POSIXct(labevents$Calender_Time, tz="UTC")
```

```
#Reading in procedures_icd.csv.gz and d_icd_procedures.csv.gz
procedures_icd <- read_csv("~/mimic/hosp/procedures_icd.csv.gz") %>%
   filter(subject_id %in% ID)
```

```
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.
```

```
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.
```

```
#Left Joining procedures_icd and d_icd_procedures by icd_code to get
#the procedure name. We only want the chartdate and long_title.
joined_procedures <- left_join(procedures_icd, d_icd_procedures,
                               by = "icd_code") %>%
  select(all_of(c("chartdate", "long_title"))) %>%
  rename("Calender_Time" = chartdate) %>%
  mutate(type = "Lab")

#Reading in transfers.csv.gz
transfers <- read_csv("~/mimic/hosp/transfers.csv.gz") %>%
  filter(subject_id %in% ID) %>%
  #Filtering out values where care units is UNKNOWN
  filter(careunit != "UNKNOWN") %>%
  select(all_of(c("subject_id", "intime", "outtime", "careunit")))
```

```
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.
```

```r
#Converting time to UTC and not PDT
transfers$intime <- as.POSIXct(transfers$intime, tz="UTC")
transfers$outtime <- as.POSIXct(transfers$outtime, tz="UTC")


#Filtering transfers by ICU/CCU
transfers_ICU_CCU <- transfers %>% filter(grepl('ICU|CCU', careunit)) %>%
  mutate(ICU_CCU = "Yes")
transfers_Not_ICU_CCU <- transfers %>% filter(!(grepl('ICU|CCU', careunit))) %>%
  mutate(ICU_CCU = "No")

#Joining resulting dataframes into 1
res_df <- bind_rows(
  labevents %>% mutate(type = "Lab"),
  joined_procedures %>% mutate(type = "Procedure"),
  transfers_ICU_CCU %>% mutate(type = "ADT"),
  transfers_Not_ICU_CCU %>% mutate(type = "ADT"),
)

#Reading in admissions.csv.gz to get race information
admissions <- read_csv("~/mimic/hosp/admissions.csv.gz") %>%
  filter(subject_id %in% ID)
```

```
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.
```

```r
subject_race <- tolower(unique(admissions$race))

#Reading in patients.csv.gz to get gender and age information
patients <- read_csv("~/mimic/hosp/patients.csv.gz") %>%
  filter(subject_id %in% ID)
```

```
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.
```

```r
subject_gender <- unique(patients$gender)
subject_age <- unique(patients$anchor_age)

subject_title <- paste0("Patient ", ID, ", ", subject_gender, ", ", subject_age,
  ", years old, ", subject_race)

#Reading in and left joining diagnoses_icd.csv.gz and d_icd_diagnoses.csv.gz
#to get the top three diagnoses for the subtitle

diagnoses_icd <- read_csv("~/mimic/hosp/diagnoses_icd.csv.gz") %>%
  filter(subject_id %in% ID) %>% head(3)
```

```
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.
```

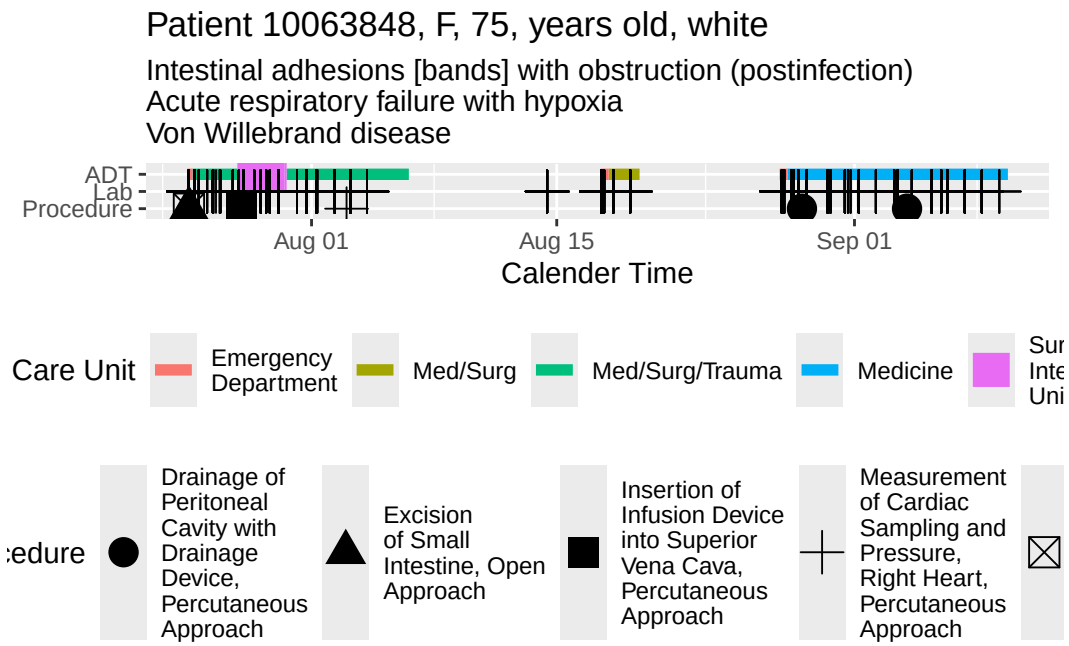
```r
d_icd_diagnoses <- read_csv("~/mimic/hosp/d_icd_diagnoses.csv.gz")
```

```
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.
```

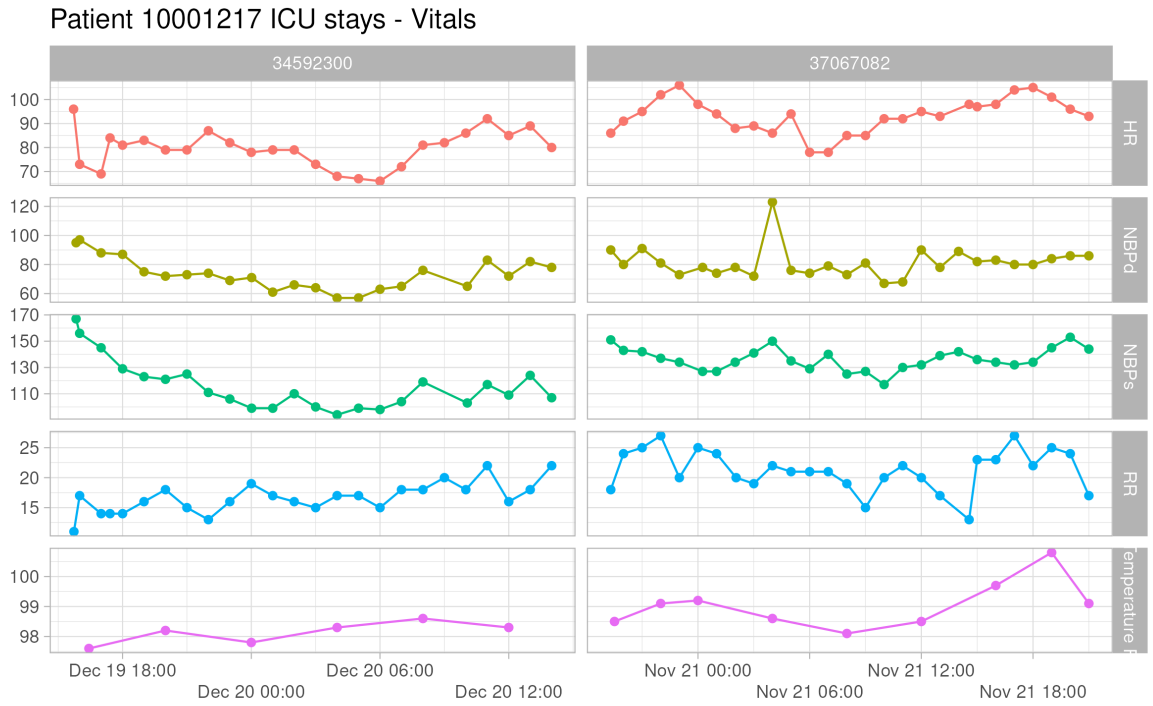
```
joined_diagnoses <- left_join(diagnoses_icd, d_icd_diagnoses,
  by = "icd_code")
top_three_diagnoses <- joined_diagnoses$long_title
subject_subtitle <- paste0(top_three_diagnoses[1], "\n",
                           top_three_diagnoses[2], "\n",
                           top_three_diagnoses[3])

#We wrap the text for our legend
res_df <- res_df %>%
  mutate(long_title = str_wrap(long_title, 15)) %>%
  mutate(careunit = str_wrap(careunit, 15))

#Define line segment size
segment_size = 2

ggplot() +
  #Plotting the ADT Data (Not ICU/CCU)
  geom_segment(data = res_df %>% filter((type == "ADT") & (ICU_CCU == "No")),
               aes(x = intime, y = type, xend = outtime, yend = type,
                   color = careunit), linewidth = segment_size) +
  #Plotting the ADT Data (ICU/CCU)
  geom_segment(data = res_df %>% filter((type == "ADT") & (ICU_CCU == "Yes")),
               aes(x = intime, y = type, xend = outtime, yend = type,
                   color = careunit), linewidth = segment_size*3) +
  #Plotting the Lab data
  geom_point(data = res_df %>% filter(type == "Lab"),
               aes(x = Calender_Time, y = type), shape = 3, size = 5) +
  #Plotting the Procedure data
  geom_point(data = res_df %>% filter(type == "Procedure"),
               aes(x = Calender_Time, y = type, shape = long_title), size = 5) +
  theme(
    axis.title.y = element_blank(),
    legend.position = "bottom",
    legend.box = "vertical"
```

```
  ) +
  scale_y_discrete(limits = rev) +
  xlab("Calender Time") +
  labs(title = subject_title,
       subtitle = subject_subtitle,
       color = "Care Unit",
       shape = "Procedure") +
  guides(color = guide_legend(order=1),
         shape = guide_legend(order=2))
```



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:

```
#Setting the subject ID
#ID <- (10001217) #This is the exmaple subject ID
ID <- (10063848)

#Reading in icustays.csv.gz
icustays <- read_csv("~/mimic/icu/icustays.csv.gz") %>%
   filter(subject_id %in% ID)
```

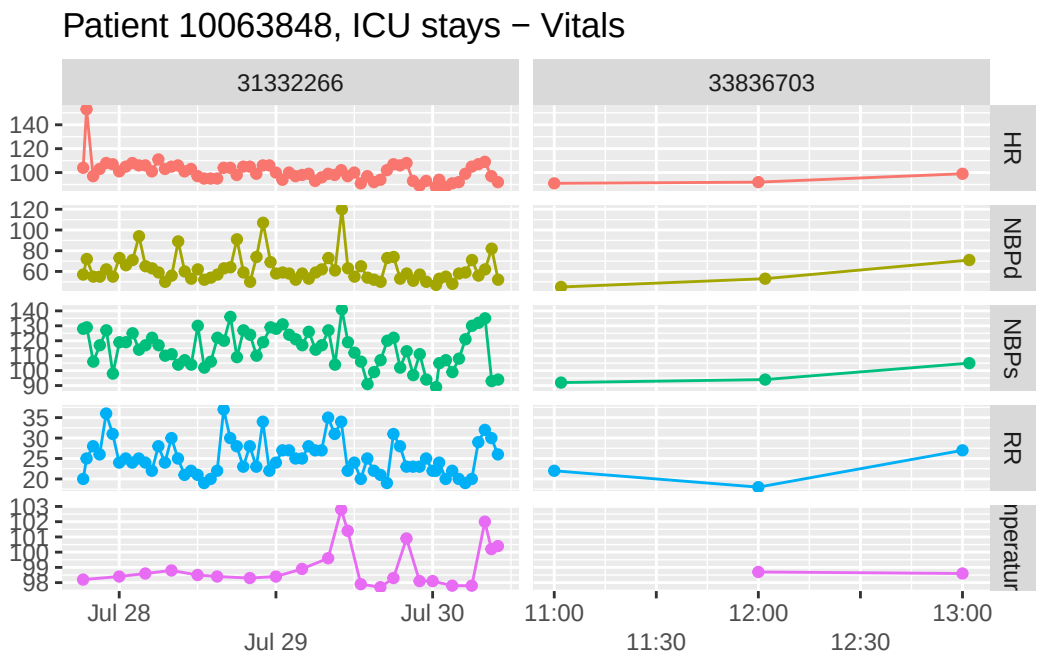
```
Rows: 94458 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (2): first_careunit, last_careunit
dbl  (4): subject_id, hadm_id, stay_id, los
dttm (2): intime, outtime

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.
```

```r
#Reading in chartevents.csv.gz, using the Parquet generated from HW2
#We create a symoblic link to the Parquet file in chartevents_pq
#We take the measurements and columns we need:
#heart rate (220045), diastolic non-invasive blood pressure (220180),
#systolic non-invasive blood pressure (220179), body temperature in
#Fahrenheit (223761), and respiratory rate (220210)

subset_itemid <-c(220045, 220180, 220179, 223761, 220210)

chartevents <- arrow::open_dataset("~/chartevents_pq") %>%
  #Filering based on subject ID
  filter(subject_id %in% ID) %>%
  #Subset based on needed measurements
  filter(itemid %in% subset_itemid) %>%
  #Sorting our values
  arrange(itemid, charttime) %>%
  collect()
#Converting time to UTC and not PDT
chartevents$charttime <- as.POSIXct(chartevents$charttime, tz="UTC")


#We left join chartevents and icustays by stay_id to differentiate the unique
#icu stays for each measurement
joined_icu_chart <- left_join(chartevents, icustays, by = "stay_id")

#We rename the item_id with their abbreviations by utilizing d_items.csv.gz
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
#We left join with our chart events on itemid
joined_icu_chart_id <- left_join(joined_icu_chart, d_items, by = "itemid")
```

```r
#Creating our title
subject_title <- paste0("Patient ", ID, ", ICU stays - Vitals")

#Note, our y is value and not valuenum, we want the values to be a double
#and not a string
ggplot(joined_icu_chart_id, mapping = aes(x = charttime, y = valuenum,
  color = abbreviation, group = abbreviation)) +
  geom_point() +
  geom_line() +
  facet_grid(abbreviation~stay_id, scales="free", space="fixed") +
  scale_x_datetime(guide = guide_axis(n.dodge = 2)) +
  theme(
    axis.title.y = element_blank(),
    axis.title.x = element_blank(),
    legend.position = "none"
  ) +
  labs(title = subject_title)
```



Q2. ICU stays

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

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

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

Q2.1 Ingestion

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

Solution:

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

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

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

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:

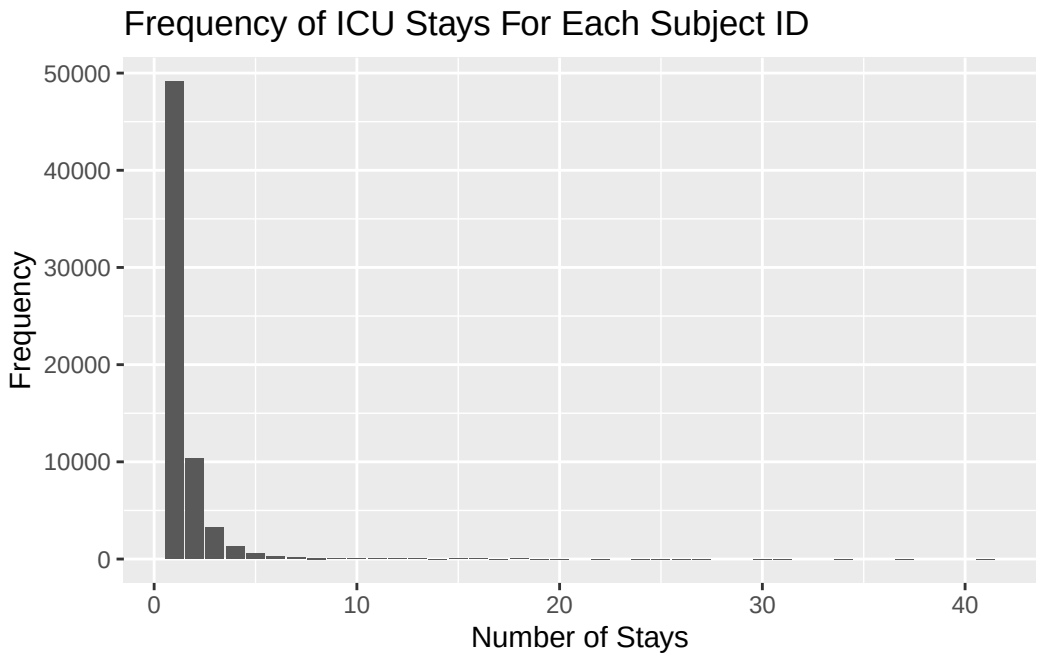
```
#Printing number of ICU stays
nrow(icustays_tble)
```

```
[1] 94458
```

```
#Printing number of unqiue subject_id
length(unique(icustays_tble$subject_id))
```

```
[1] 65366
```

```
num_ICU_stays_per_subject_id <- icustays_tble %>%
  #Group by subject_id
  group_by(subject_id) %>%
  #Count number of stays per subject_id
  summarise(n_stays = n()) %>%
  #Group by number of stays
  group_by(n_stays) %>%
  #Count the number of times for each number of stays
  summarise(freq_n_stays = n())

ggplot(num_ICU_stays_per_subject_id, aes(x = n_stays,
  y = freq_n_stays)) +
  geom_bar(stat = 'identity') +
  xlab("Number of Stays") +
  ylab("Frequency") +
  theme(legend.position = "none") +
  labs(title = "Frequency of ICU Stays For Each Subject ID")
```



The number of unique `subject_id` is 65366 while the total number of ICU stays is 94458, which means that there are some subjects who have multiple ICU stays.

Q3. `admissions` data

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

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

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

Q3.1 Ingestion

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

Solution:

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

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

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

Q3.2 Summary and visualization

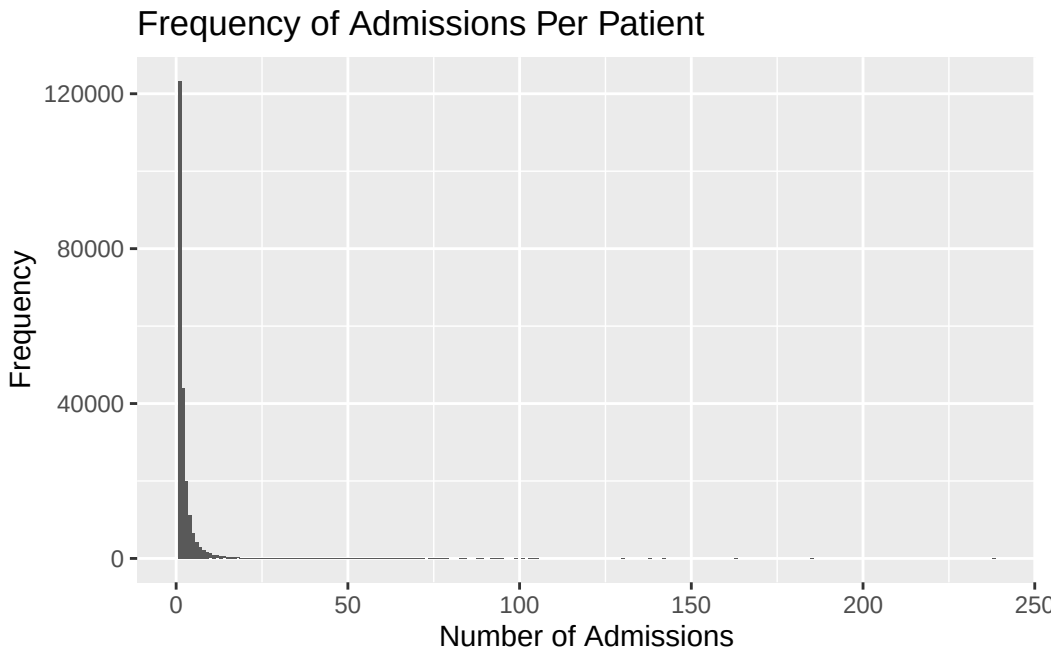
Summarize the following information by graphics and explain any patterns you see.

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

Solution:

```
#Number of admissions per patient
num_admissions_per_patient <- admissions_tble %>%
  #Group by subject_id
  group_by(subject_id) %>%
  #Count number of admissions per patient
  summarise(n_admissions = n()) %>%
  #Group by number of admissions
  group_by(n_admissions) %>%
  #Count the number of times for each number of stays
  summarise(freq_n_admissions = n())
```

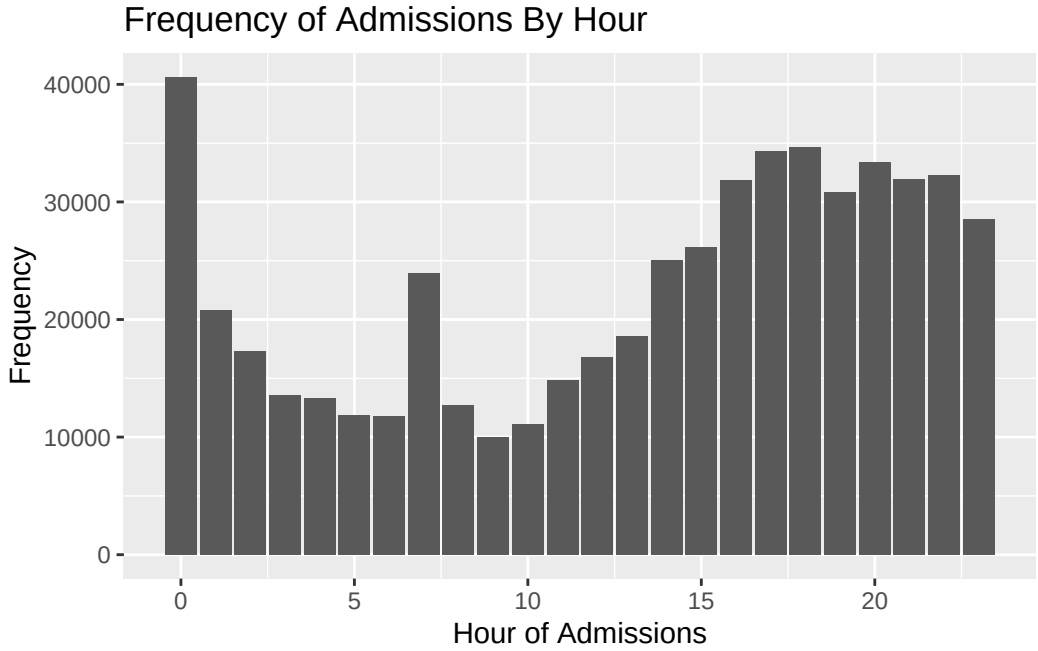
```
ggplot(num_admissions_per_patient, aes(x = n_admissions,
  y = freq_n_admissions)) +
  geom_bar(stat = 'identity') +
  xlab("Number of Admissions") +
  ylab("Frequency") +
  theme(legend.position = "none") +
  labs(title = "Frequency of Admissions Per Patient")
```



The chart demonstrates most patients have less than ten hospital admissions because most patients have their conditions treated/cured after a few hospital visits. However, we see that there exist some outliers that have more than a hundred hospital admissions. However, we do not know if this is an outlier due to a technical error or if this patient is chronically sick and requires constant hospital care.

```
#Admission hour

#Adding an hour column to admissions_tble

admissions_tble$admission_hour <- hour(admissions_tble$admittime)

n_hour_of_admission <- admissions_tble %>%
  #Group by admission hour
```

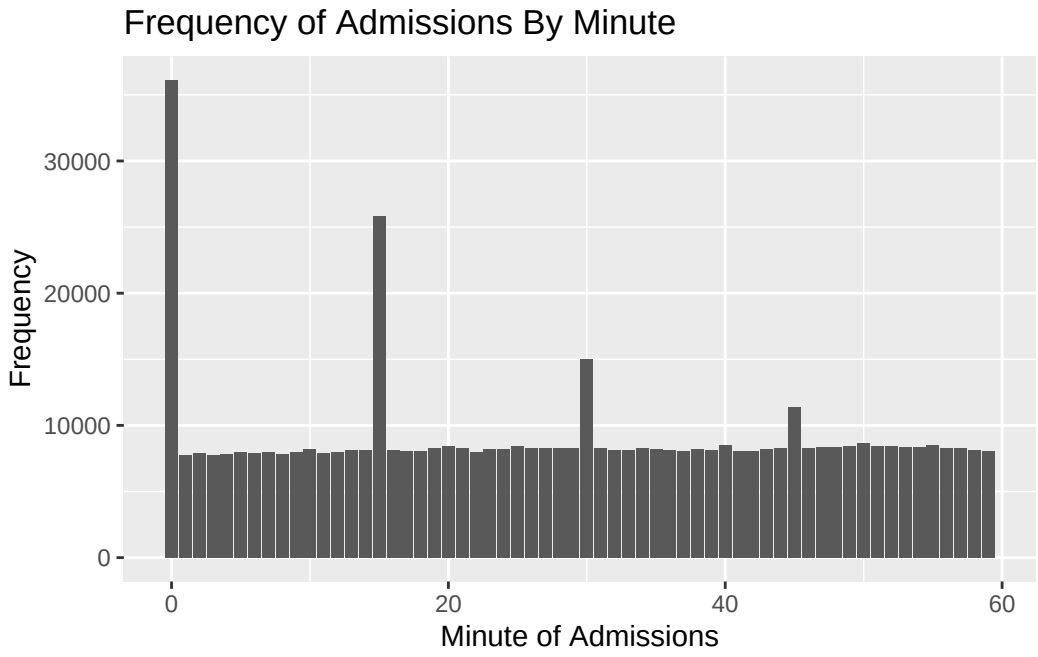
```
  group_by(admission_hour) %>%
  #Count number of admissions by hour
  summarise(n_admissions_hour = n())

ggplot(n_hour_of_admission, aes(x = admission_hour,
  y = n_admissions_hour)) +
  geom_bar(stat = 'identity') +
  xlab("Hour of Admissions") +
  ylab("Frequency") +
  theme(legend.position = "none") +
  labs(title = "Frequency of Admissions By Hour")
```



The graph demonstrates that most hospital admissions occur at around 7 a.m. or in the afternoon/nighttime from 3 p.m. to midnight. This small initial spike in admissions is likely because most people try to get hospital appointments before work/school at 8 a.m. or 9 a.m. As people get off work or school around 3 p.m., they can make it to the hospital.

```
#Admission Minute

#Adding a minute column to admissions_tble

admissions_tble$admission_minute <- minute(admissions_tble$admittime)
```

```r
n_minute_of_admission <- admissions_tble %>%
  #Group by admission minute
  group_by(admission_minute) %>%
  #Count number of admissions by minute
  summarise(n_admissions_minute = n())

ggplot(n_minute_of_admission, aes(x = admission_minute,
  y = n_admissions_minute)) +
  geom_bar(stat = 'identity') +
  xlab("Minute of Admissions") +
  ylab("Frequency") +
  theme(legend.position = "none") +
  labs(title = "Frequency of Admissions By Minute")
```

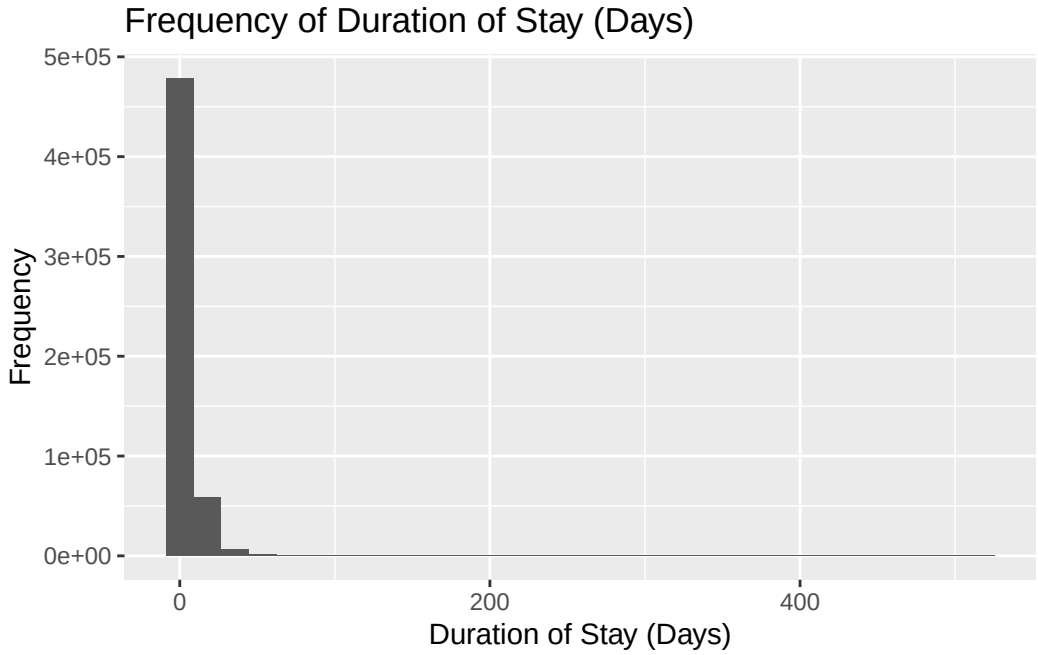


This graph demonstrates that most hospital admissions occur at 15-minute intervals starting at the top of the hour. A couple of possible explanations for this pattern are that hospital appointments tend to last in intervals of 15 minutes, from half an hour to 45 minutes to one hour, and so on. This results in appointment admission times scheduled around these intervals.

```r
#Length of Hospital Stay

#Find the duration of the length of stay
```

```r
admissions_tble$length_of_stay <- difftime(admissions_tble$dischtime,
  admissions_tble$admittime, units="days")

ggplot(admissions_tble, aes(x = length_of_stay)) +
  geom_histogram() +
  stat_bin(bins = 30) +
  xlab("Duration of Stay (Days)") +
  ylab("Frequency") +
  theme(legend.position = "none") +
  labs(title = "Frequency of Duration of Stay (Days)")
```

```
Don't know how to automatically pick scale for object of type <difftime>.
Defaulting to continuous.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



This graph demonstrates that the typical duration of stays at the hospital does not last more than 10 days. This pattern means most ailments are treated within that period, except for severe conditions requiring constant and longstanding treatment plans. There are some outliers in stay duration lasting over a year, though this could be an entry error or a patient with a condition that needs lengthy care, such as a coma.

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.

Q4. `patients` data

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

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

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

Q4.1 Ingestion

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

Solution:

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

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

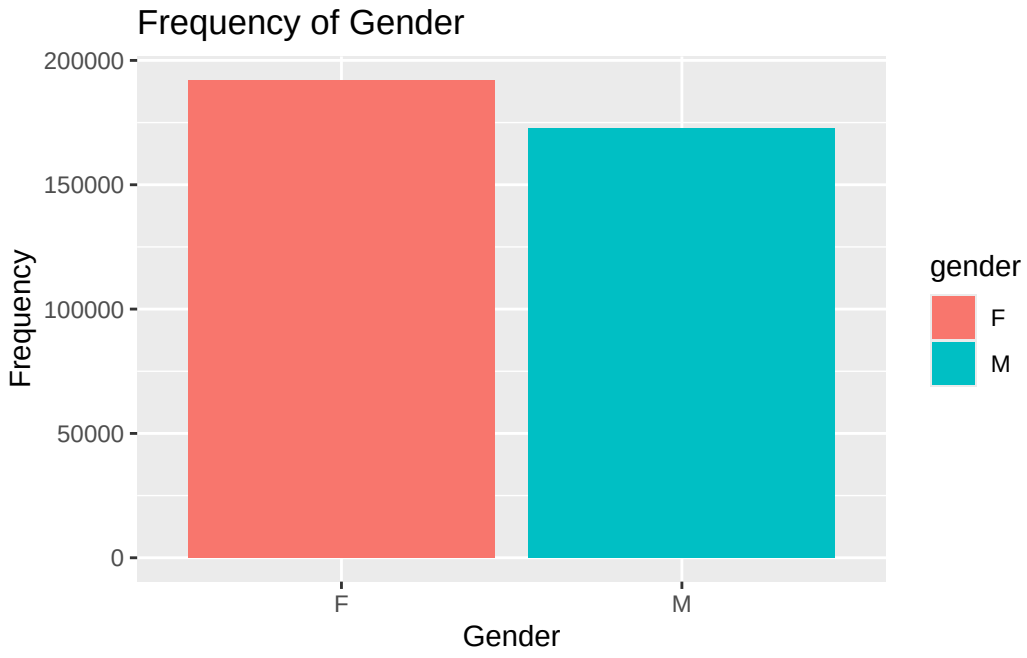
```
i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Q4.2 Summary and visualization

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

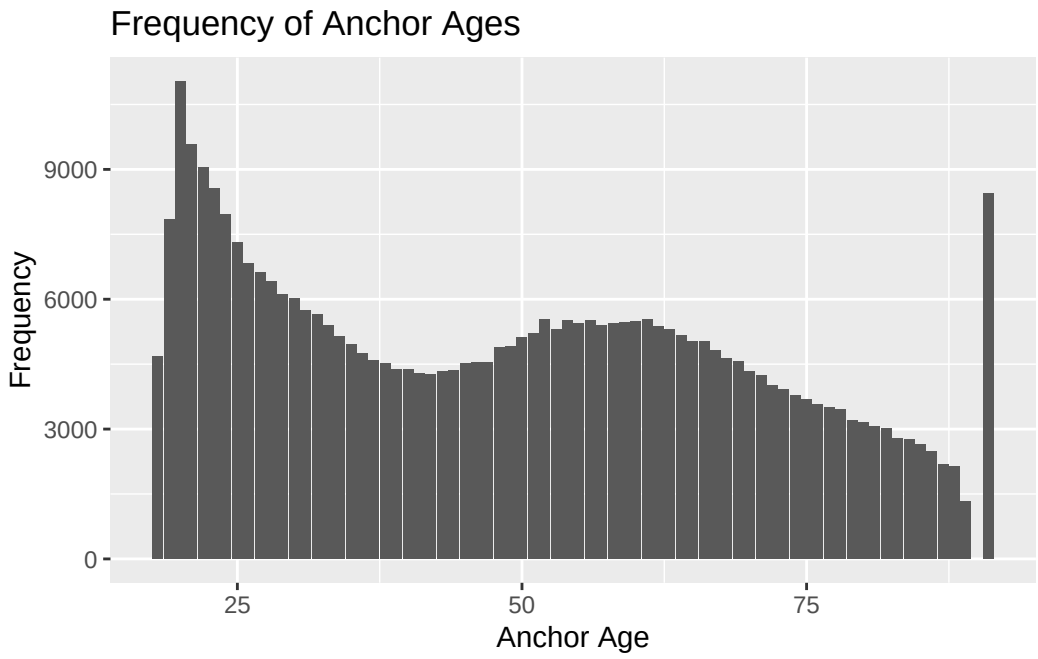
Solution:

```r
#Summarizing Gender
n_gender <- patients_tble %>%
  #Group by gender
  group_by(gender) %>%
  #Count number of each gender
  summarise(n_gender = n())

ggplot(n_gender, aes(x = gender, y = n_gender, fill = gender)) +
  geom_bar(stat = 'identity') +
  xlab("Gender") +
  ylab("Frequency") +
  theme(legend.position = "right") +
  labs(title = "Frequency of Gender")
```



There is not much of a pattern for patient frequencies by gender, except for a slightly higher frequency of female patients. This pattern could be the result of male patients' reluctance to go to a healthcare facility and instead opting to suck it up.

```
#Summarizing anchor_age
n_anchor_age <- patients_tble %>%
  #Group by anchor_age
  group_by(anchor_age) %>%
  #Count number of anchor_age
  summarise(n_anchor_age = n())

ggplot(n_anchor_age, aes(x = anchor_age, y = n_anchor_age)) +
  geom_bar(stat = 'identity') +
  xlab("Anchor Age") +
  ylab("Frequency") +
  theme(legend.position = "none") +
  labs(title = "Frequency of Anchor Ages")
```



This chart demonstrates a clear spike in patients' ages around the early teens and preteens. This pattern can be explained by children being more susceptible to becoming sick or injuring themselves and cautious parents feeling the need to admit their children to be safe. There is a decline in patients from 35-45, as health problems do not appear again until patients reach around 50 years of age. As patients pass away of old age, fewer patients are admitted in the later age ranges.

Q5. Lab results

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

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

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

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

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

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

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

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

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

Solution:

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

#We read in d_labitems by to join with labevents_tble later to get the labels
#for the items

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

```
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 <- arrow::open_dataset("~/labevents_pq") %>%
  #Converting to duckdb
  arrow::to_duckdb() %>%
  #Filtering by measurements we want to take
  filter(itemid %in% subset_itemid) %>%
  #We keep only the columns we need
  select(subject_id, itemid, storetime, valuenum) %>%
```

```
  #Mutating columns for proper joins
  #Mutate subject_id and hadm_id as double to join with icustays_tble
  mutate(subject_id = as.double(subject_id)) %>%
  #We inner join with icu stays, keeping only the patients with an icu stay
  #We inner join to discard nonmatching rows from both tibbles
  inner_join(icustays_tble, by = c("subject_id"), copy = TRUE) %>%
  filter(storetime < intime) %>%
  #We group by stay_id and itemid to get the values for each measurement
  #of each stay
  group_by(subject_id, stay_id, itemid) %>%
  #We order by the store time, taking the last measurement before intime
  #Using slice_max(), we take a slice of size one for the highest time
  slice_max(order_by = storetime, n = 1) %>%
  summarize(valuenum = mean(valuenum, na.rm = TRUE)) %>%
  #We ungroup to get the dataframe back to a normal size
  ungroup() %>%
  #We join labevents_tble with d_labitems by itemid
  left_join(d_labitems, by = "itemid", copy = TRUE) %>%
  #We subset labevents_tbl to only the columns we need in our final result
  select(c(subject_id, stay_id, valuenum, label)) %>%
  #Apply lower case to all labels
  mutate(label = tolower(label)) %>%
  #We widen the dataframe to get each row as a subject and ICU stay
  pivot_wider(names_from = label, values_from = valuenum) %>%
  #Sorting the tble by subject_id and_stay id for grading purposes
  arrange(subject_id, stay_id) %>%
  #Changing white blood cells to wbc, removing spaces
  rename(wbc = `white blood cells`) %>%
  collect()
```

```
`summarise()` has grouped output by "subject_id" and "stay_id". You can
override using the `.groups` argument.
`summarise()` has grouped output by "subject_id" and "stay_id". You can
override using the `.groups` argument.
```

```
labevents_tble
```

```
# A tibble: 88,086 x 10
   subject_id  stay_id potassium   wbc bicarbonate chloride glucose creatinine
        <dbl>    <dbl>     <dbl> <dbl>       <dbl>    <dbl>   <dbl>      <dbl>
 1   10000032 39553978       6.7   6.9          25       95     102        0.7
```

```
 2   10000690 37081114       4.8   7.1          26       100       85        1
 3   10000980 39765666       3.9   5.3          21       109       89        2.3
 4   10001217 34592300       4.1   5.4          30       104       87        0.5
 5   10001217 37067082       4.2  15.7          22       108      112        0.6
 6   10001725 31205490       4.1  NA            NA        98       NA       NA
 7   10001843 39698942       3.9  10.4          28        97      131        1.3
 8   10001884 37510196       4.5  12.2          30        88      141        1.1
 9   10002013 39060235       3.5   7.2          24       102      288        0.9
10   10002114 34672098       6.5  16.8          18        NA       95        3.1
# i 88,076 more rows
# i 2 more variables: sodium <dbl>, hematocrit <dbl>
```

```
#Clearing up memory for next step
gc()
```

```
           used  (Mb) gc trigger  (Mb) max used  (Mb)
Ncells  2586556 138.2    5020434 268.2  5020434 268.2
Vcells 21085152 160.9   48717048 371.7 60896305 464.7
```

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.

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

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

Solution:

```
subset_itemid <-c(220045, 220180, 220179, 223761, 220210)

#We read in d_items to join with chartevents_tble and get the labels
#for the items later
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_tble <- arrow::open_dataset("~/chartevents_pq") %>%
  #Converting to duckdb
  arrow::to_duckdb() %>%
  #Subset based on needed measurements
  filter(itemid %in% subset_itemid) %>%
  #We keep only the columns we need
  select(subject_id, stay_id, itemid, storetime, valuenum) %>%
  #Mutate subject_id and stay_id as double to join with icustays_tble
  mutate(subject_id = as.double(subject_id)) %>%
  mutate(stay_id = as.double(stay_id)) %>%
  #Inner join with ICU Stays
  #Keep patients with an ICU stay
  inner_join(icustays_tble, by = c("subject_id", "stay_id"), copy = TRUE) %>%
  #Filter by measurements within the ICU stay
  filter((storetime > intime) & (storetime < outtime)) %>%
  #We group by stay_id and itemid to get the values for each measurement
  #of each stay
  group_by(subject_id, stay_id, itemid) %>%
  #We order by the store time, taking the first measurement during the ICU stay
  #Using slice_min(), we take a slice of size one for the smallest time in that
  #interval
  slice_min(order_by = storetime, n = 1) %>%
  summarize(valuenum = mean(valuenum, na.rm = TRUE)) %>%
  #We ungroup to get the dataframe back to a normal size
  ungroup() %>%
  left_join(d_items, by = "itemid", copy = TRUE) %>%
  select(c(subject_id, stay_id, valuenum, label)) %>%
  #Apply lower case to all labels and remove sapce
  mutate(label = str_replace_all(tolower(label), " ", "_")) %>%
  #We widen the dataframe to get each row as a subject and ICU stay
  pivot_wider(names_from = label, values_from = valuenum) %>%
  #Sorting the tble by subject_id and_stay id for grading purposes
  arrange(subject_id, stay_id) %>%
```

```
  collect()
```

```
`summarise()` has grouped output by "subject_id" and "stay_id". You can
override using the `.groups` argument.
`summarise()` has grouped output by "subject_id" and "stay_id". You can
override using the `.groups` argument.
```

```
chartevents_tble
```

```
# A tibble: 94,363 x 7
   subject_id  stay_id respiratory_rate non_invasive_blood_pressure_systolic
        <dbl>    <dbl>            <dbl>                                <dbl>
 1   10000032 39553978             24                                   84
 2   10000690 37081114             24.3                                106
 3   10000980 39765666             23.5                                154
 4   10001217 34592300             14                                  156
 5   10001217 37067082             18                                  151
 6   10001725 31205490             19                                   73
 7   10001843 39698942             16.5                                110
 8   10001884 37510196             13                                  174.
 9   10002013 39060235             14                                   98.5
10   10002114 34672098             21                                  112
# i 94,353 more rows
# i 3 more variables: non_invasive_blood_pressure_diastolic <dbl>,
#   temperature_fahrenheit <dbl>, heart_rate <dbl>
```

```
#Clearing up memory for next step
gc()
```

```
          used  (Mb) gc trigger  (Mb) max used  (Mb)
Ncells  2586795 138.2    5020434 268.2  5020434 268.2
Vcells 21746037 166.0   48717048 371.7 60896305 464.7
```

Q7. Putting things together

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

- all variables in `icustays_tble`

- all variables in `admissions_tble`

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

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

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

Solution:

```
mimic_icu_cohort <- icustays_tble %>%
  #Left join with admissions_tble by subject_id and hadm_id
  left_join(admissions_tble, by = c("subject_id", "hadm_id")) %>%
  #Left join with patients_tbl by subject_id
  left_join(patients_tble, by = c("subject_id")) %>%
  #Left join with labevents_tble by subject_id and stay_id
  left_join(labevents_tble, by = c("subject_id", "stay_id")) %>%
  #Left join with chartevents_tble by subject_id and stay_id
  left_join(chartevents_tble, by = c("subject_id", "stay_id")) %>%
  #Filter by adults (anchor age >= 18)
  filter(anchor_age >= 18) %>%
  #Sorting the tble by subject_id and_stay id for grading purposes
  arrange(subject_id, stay_id)

mimic_icu_cohort
```

```
# A tibble: 94,458 x 43
   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 27703517 34592300 Surgical Inte~ Surgical Int~ 2157-12-19 15:42:24
 5   10001217 24597018 37067082 Surgical Inte~ Surgical Int~ 2157-11-20 19:18:02
 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 37 more variables: outtime <dttm>, los <dbl>, admittime <dttm>,
#   dischtime <dttm>, deathtime <dttm>, admission_type <chr>,
#   admit_provider_id <chr>, admission_location <chr>,
#   discharge_location <chr>, insurance <chr>, language <chr>,
#   marital_status <chr>, race <chr>, edregtime <dttm>, edouttime <dttm>,
#   hospital_expire_flag <dbl>, admission_hour <int>, ...
```

Q8. Exploratory data analysis (EDA)

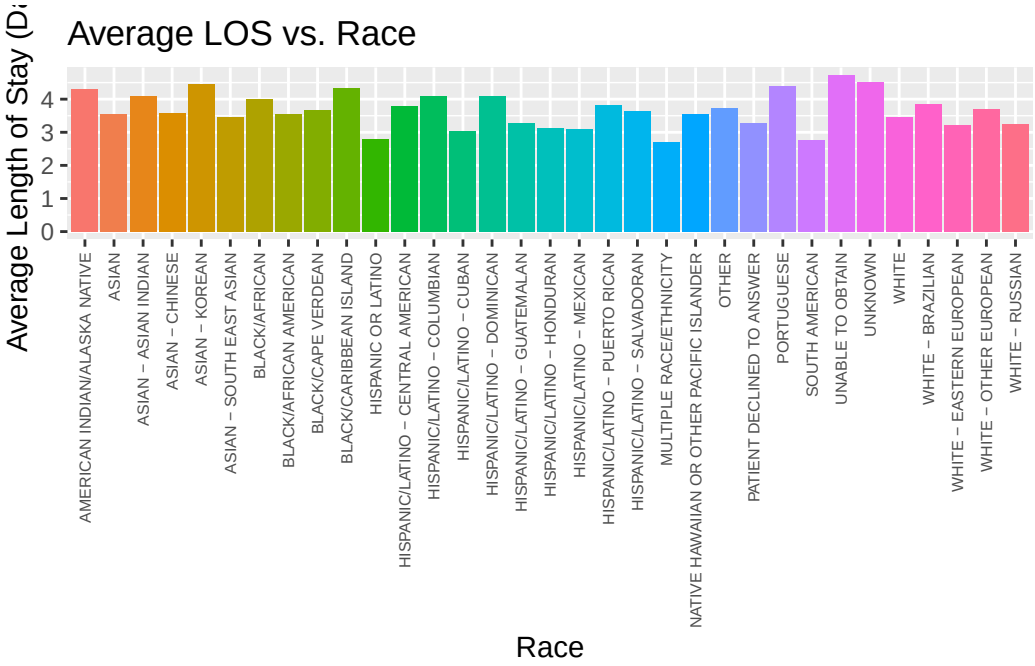
Summarize the following information about the ICU stay cohort `mimic_icu_cohort` using appropriate numerics or graphs:

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

Solution:

```
#For each demographic, we group by the demographic and take the mean of the
#length of stay.

#Race
los_vs_demo_race <- mimic_icu_cohort %>%
  group_by(race) %>%
  summarize(average_length_of_stay = mean(los, na.rm = TRUE)) %>%
  arrange(average_length_of_stay)

ggplot(los_vs_demo_race, mapping = aes(x = race, y = average_length_of_stay,
  fill = race)) +
  geom_bar(stat = "identity") +
  #Rotating the x-axis
```

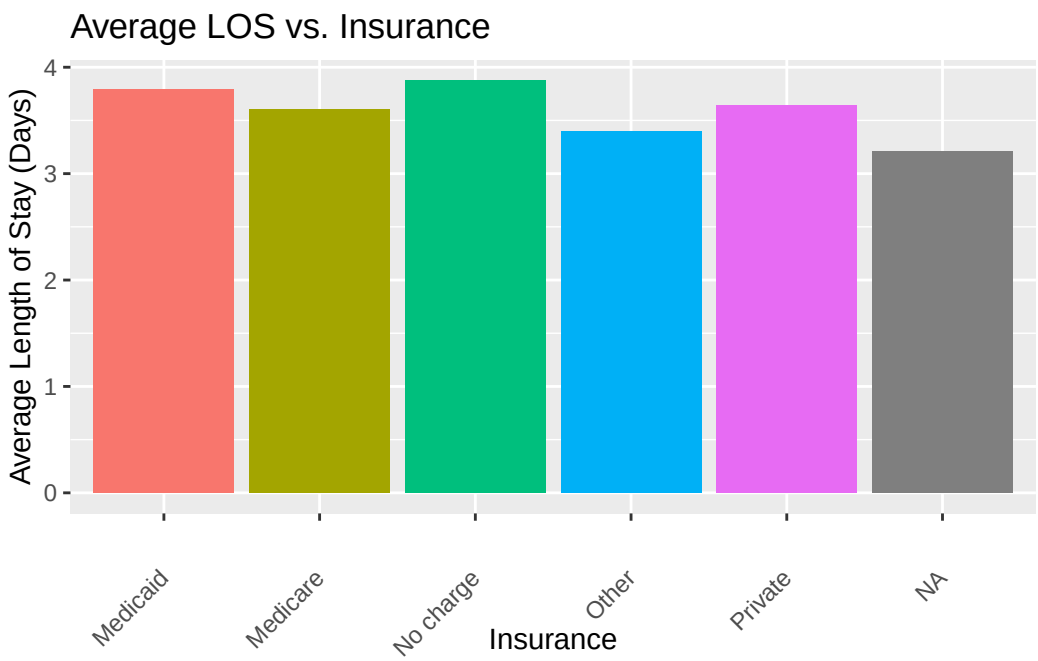
```
  theme(axis.text.x = element_text(size = 6, angle = 90, vjust = 0.5, hjust=1),
        legend.position="none") +
  ylab("Average Length of Stay (Days)") +
  xlab("Race") +
  ggtitle("Average LOS vs. Race")
```



This chart demonstrates that South American and Hispanic/Latino patients tend to stay the least amount of time in the ICU. While any explanation behind this pattern is a hypothesis, some possible explanations can be fear of medical bills or distrust of the medical systems. Unknown and unable to obtain races have the highest average length of stay. The higher average length of stay could result from patients in a very critical condition where they cannot provide this information.

```
#Insurance
los_vs_demo_insurance <- mimic_icu_cohort %>%
  group_by(insurance) %>%
  summarize(average_length_of_stay = mean(los, na.rm = TRUE)) %>%
  arrange(average_length_of_stay)

ggplot(los_vs_demo_insurance, mapping = aes(x = insurance,
  y = average_length_of_stay, fill = insurance)) +
  geom_bar(stat = "identity") +
  #Rotating the x-axis
```

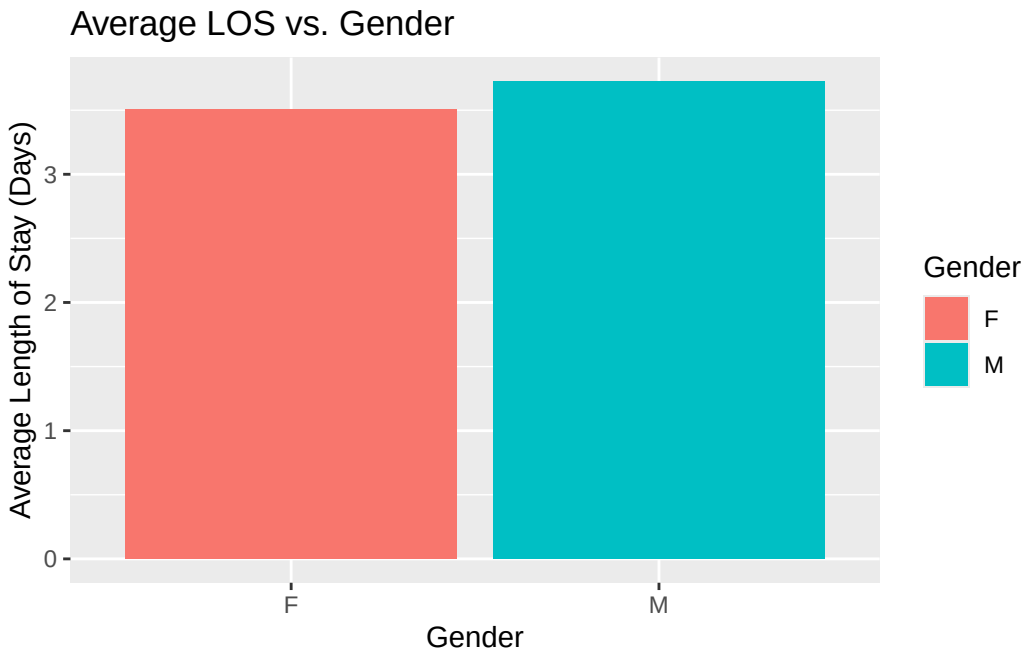
```
  theme(axis.text.x = element_text(angle = 45, vjust = 0.5, hjust=1),
        legend.position="none") +
  ylab("Average Length of Stay (Days)") +
  xlab("Insurance") +
  ggtitle("Average LOS vs. Insurance")
```



There is not much of a pattern of the average length of ICU stay depending on the type of medical insurance. Patients with no (N/A) insurance tend to stay slightly less than those covered by insurance. This pattern is likely due to patients without insurance being more cost-conscious about their stay.

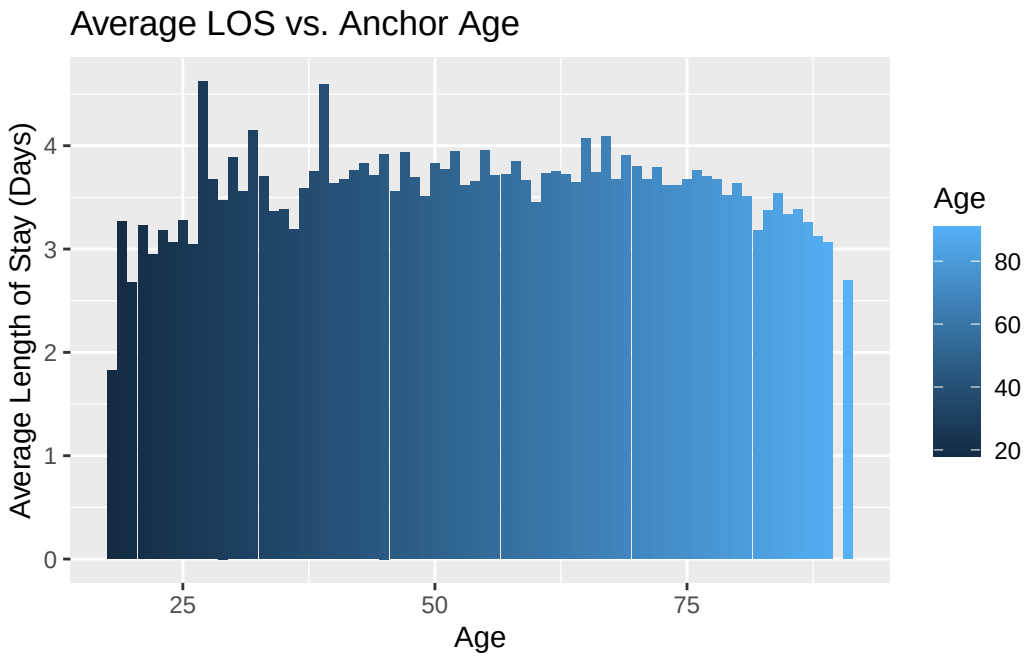
```
#Gender
los_vs_demo_gender <- mimic_icu_cohort %>%
  group_by(gender) %>%
  summarize(average_length_of_stay = mean(los, na.rm = TRUE)) %>%
  arrange(average_length_of_stay)

ggplot(los_vs_demo_gender, mapping = aes(x = gender,
  y = average_length_of_stay, fill = gender)) +
  geom_bar(stat = "identity") +
  labs(fill = "Gender") +
  ylab("Average Length of Stay (Days)") +
```

```
  xlab("Gender") +
  ggtitle("Average LOS vs. Gender")
```



There is not a large difference/noticeable pattern in the average length of ICU stay between males and females. The small difference could be the result of sampling.

```
#Age at intime
los_vs_demo_age_at_intime <- mimic_icu_cohort %>%
  group_by(anchor_age) %>%
  summarize(average_length_of_stay = mean(los, na.rm = TRUE))

ggplot(los_vs_demo_age_at_intime, mapping = aes(x = anchor_age,
  y = average_length_of_stay, fill = anchor_age)) +
  geom_bar(stat = "identity") +
  labs(fill = "Age") +
  ylab("Average Length of Stay (Days)") +
  xlab("Age") +
  ggtitle("Average LOS vs. Anchor Age")
```

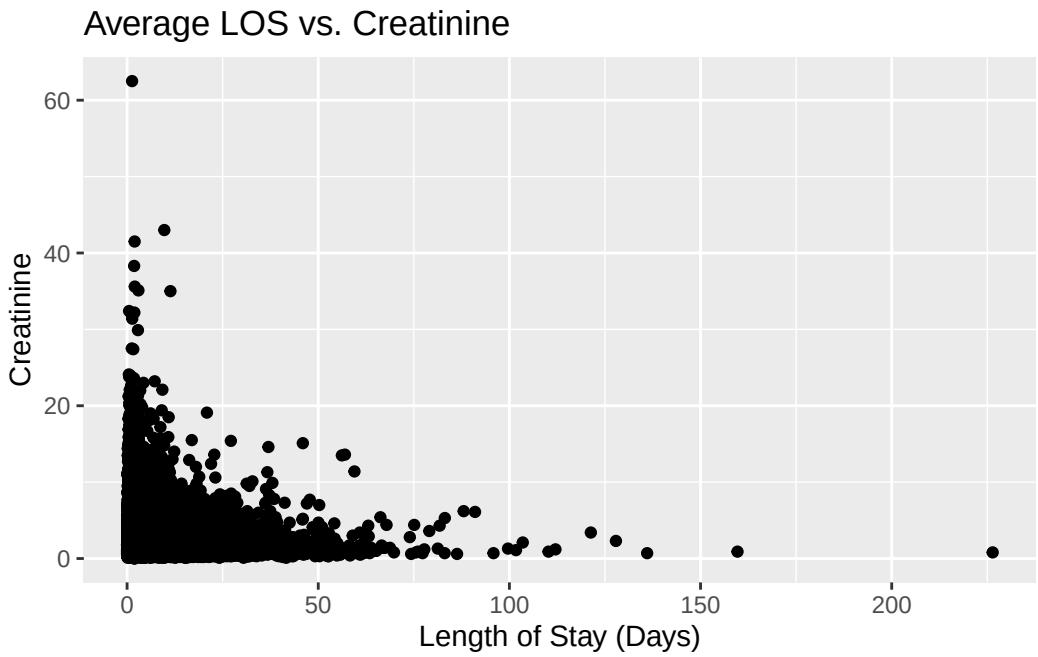


The average stay length is consistent throughout the ages except for the early teens and later in life. ICU patients in their early teens could have a faster recovery time due to their youth, and ICU patients later in life could not have much life left in them, leading to shorter ICU times before they pass away.

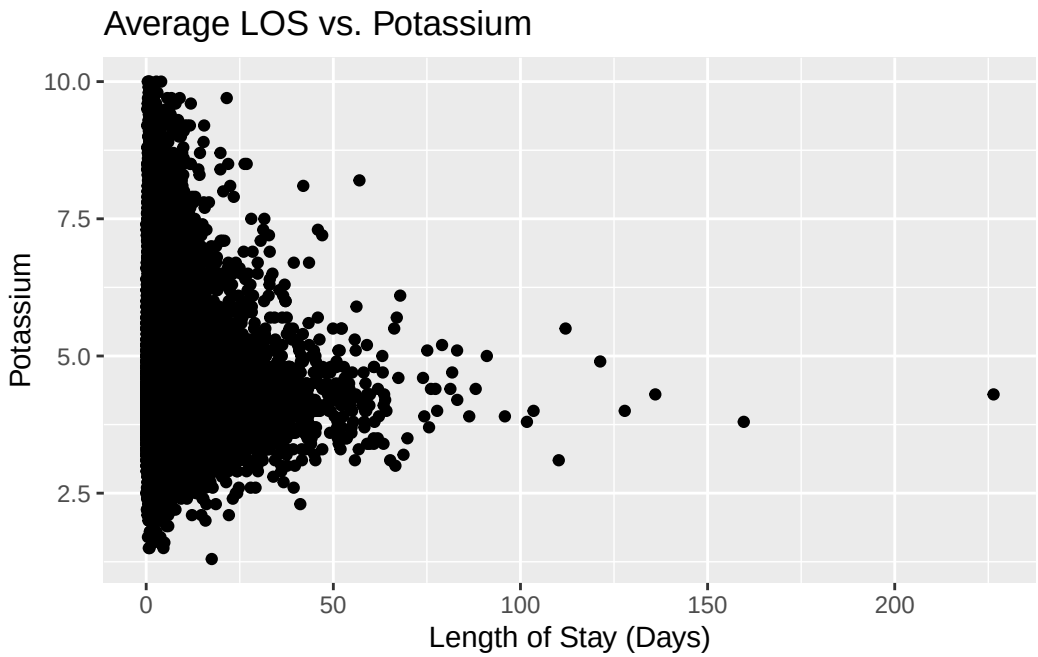
```
#Create a list of measurements taken before ICU Stay
#creatinine, potassium, sodium, chloride, bicarbonate, hematocrit,
#white blood cell count, glucose (50931)
measure_name_before_ICU_stay = c('creatinine', 'potassium', 'sodium',
  'chloride', 'bicarbonate', 'hematocrit', 'wbc', 'glucose')

for (measurement in measure_name_before_ICU_stay) {
    label_title = str_to_title(str_replace_all(measurement, "_", " "))
    print(ggplot(mimic_icu_cohort, aes(x = los,
      y = mimic_icu_cohort[[measurement]])) +
      geom_point() +
    xlab("Length of Stay (Days)") +
    ylab(label_title) +
    ggtitle(str_glue("Average LOS vs. {label_title}"))
      )
}
```

```
Warning: Removed 8041 rows containing missing values or values outside the scale range
```

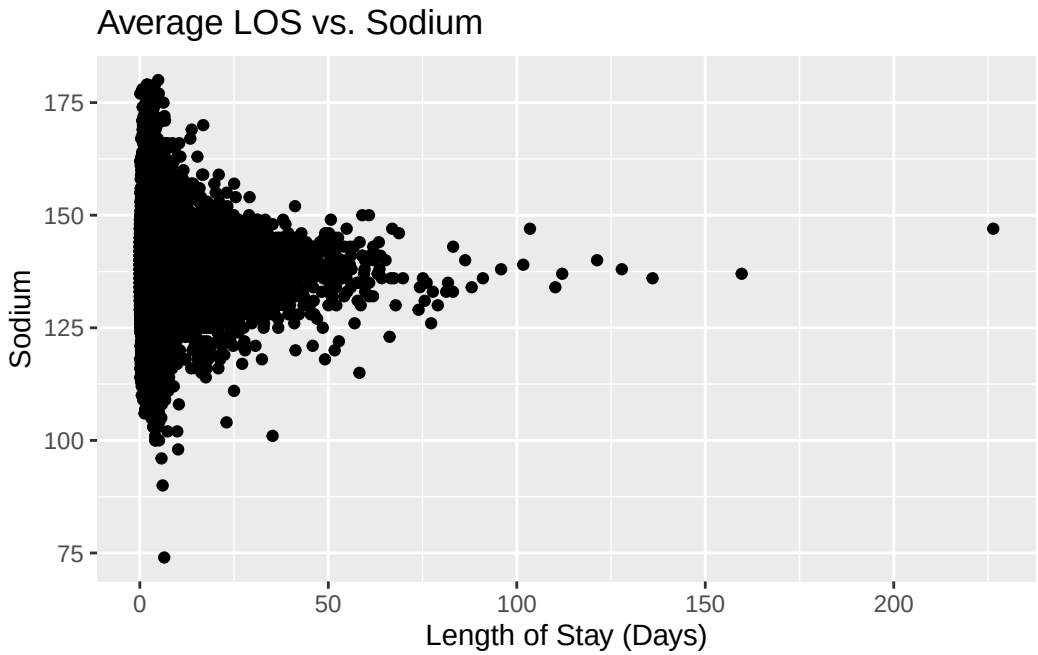
(`geom_point()`).



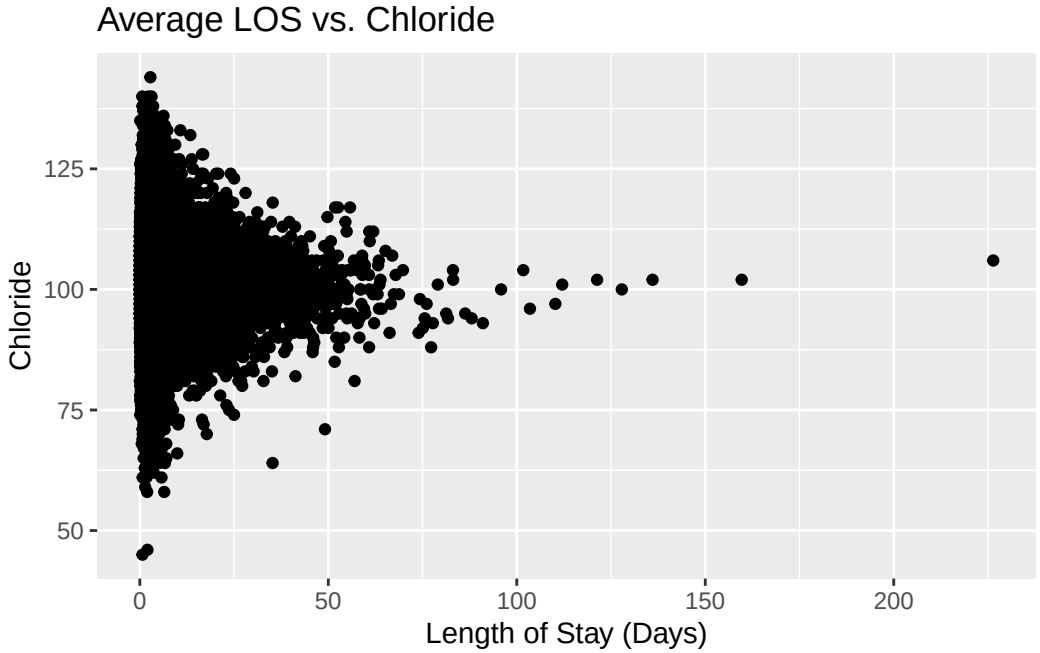
Warning: Removed 11401 rows containing missing values or values outside the scale range (`geom_point()`).



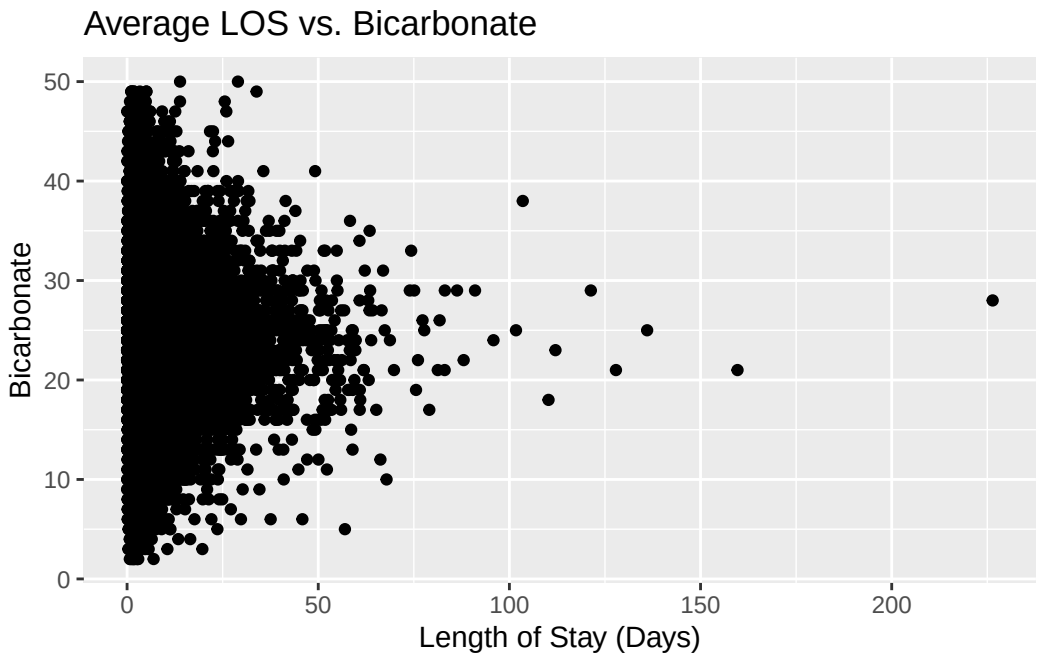

```
Warning: Removed 11344 rows containing missing values or values outside the scale range
(`geom_point()`).
```



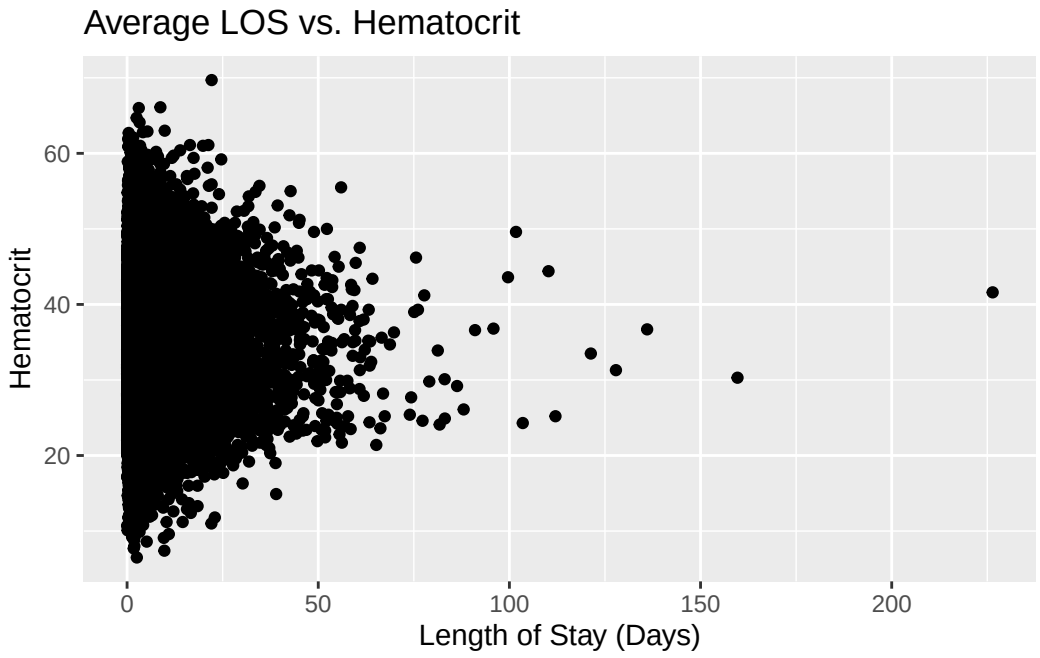
```
Warning: Removed 11365 rows containing missing values or values outside the scale range
(`geom_point()`).
```



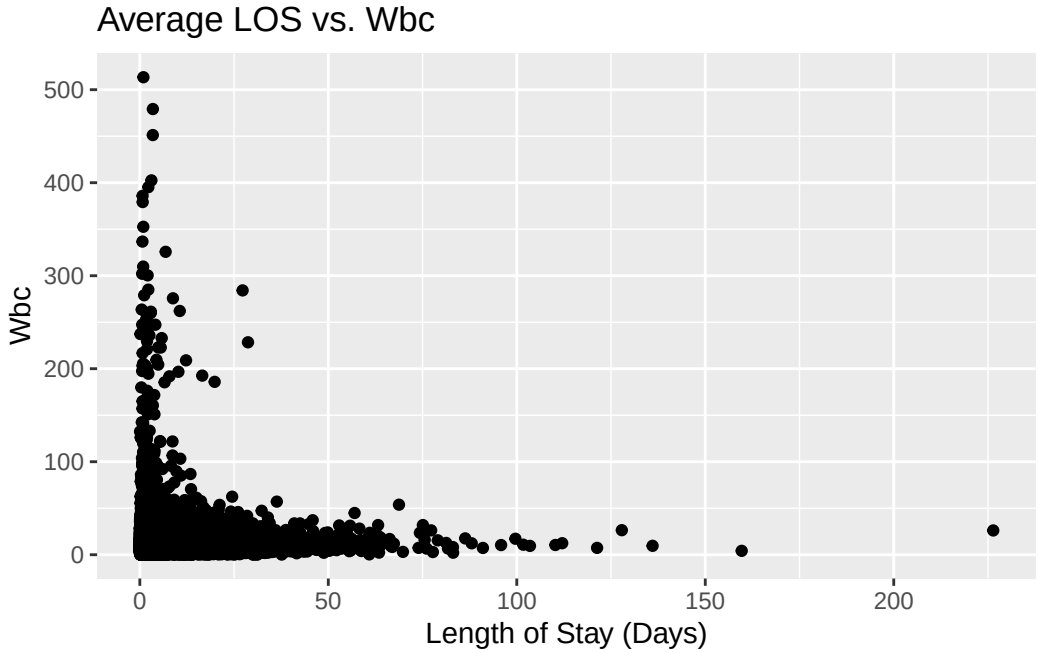

```
Warning: Removed 11563 rows containing missing values or values outside the scale range
(`geom_point()`).
```



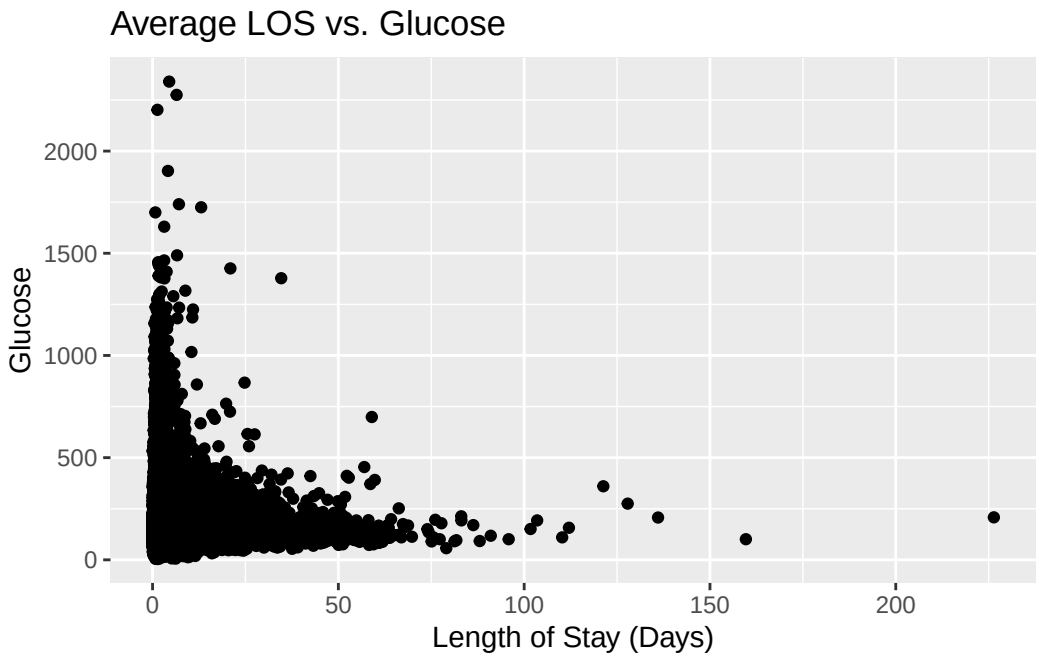

```
Warning: Removed 6765 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
Warning: Removed 6864 rows containing missing values or values outside the scale range
(`geom_point()`).
```




```
Warning: Removed 11668 rows containing missing values or values outside the scale range
(`geom_point()`).
```

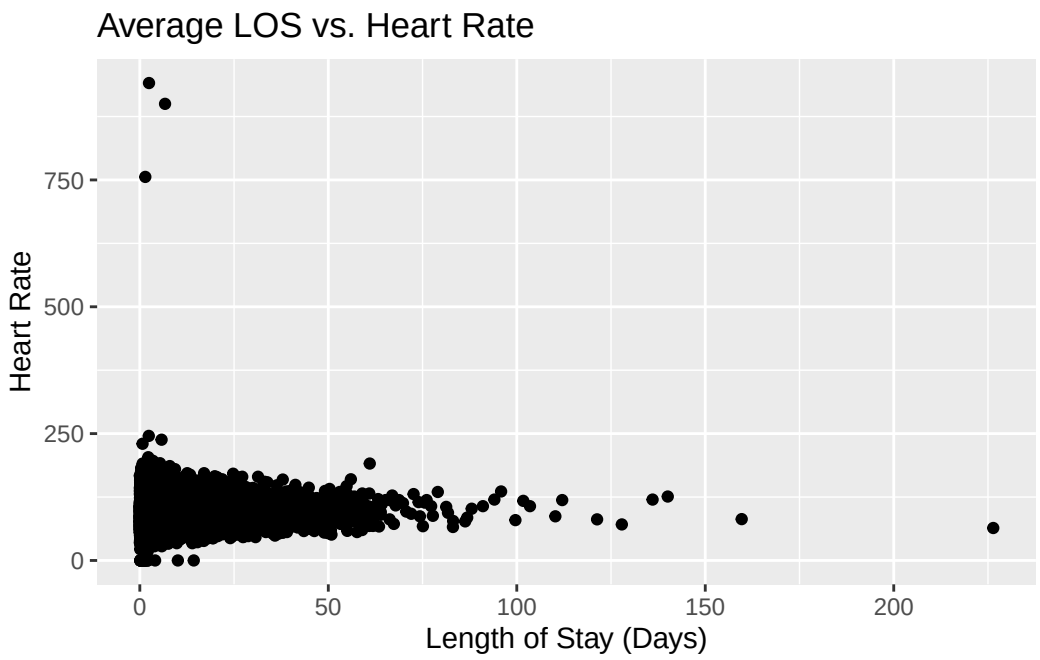


These charts demonstrate some patients have abnormally low or high last measurements before their ICU stay. This pattern could result from a measurement spike that needed short, critical care to restabilize them. For patients with longer stays, the measurements appear more stable. This stability must mean that patients with lengthier stays have a condition unrelated to these measurements.

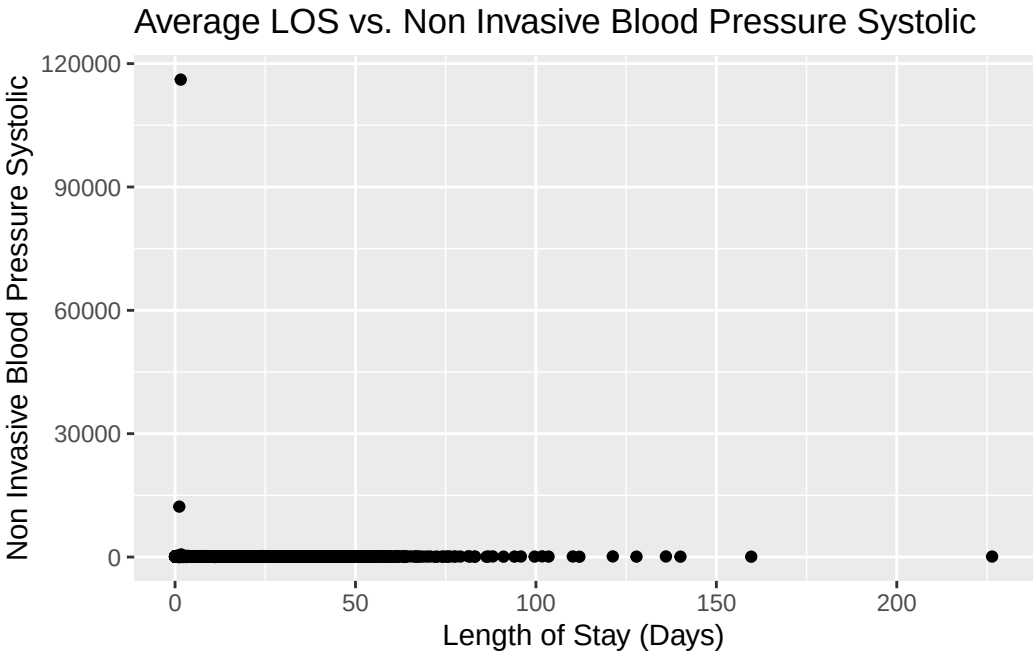
```
#Create a list of measurements taken after ICU Stay
#heart rate, systolic non-invasive blood pressure, diastolic non-invasive blood
#pressure, body temperature in Fahrenheit, and respiratory rate
measure_name_after_ICU_stay = c('heart_rate',
                                'non_invasive_blood_pressure_systolic',
                                'non_invasive_blood_pressure_diastolic',
                                'respiratory_rate',
                                'temperature_fahrenheit'
                                )
for (measurement in measure_name_after_ICU_stay) {
    label_title = str_to_title(str_replace_all(measurement, "_", " "))
    print(ggplot(mimic_icu_cohort, aes(x = los,
      y = mimic_icu_cohort[[measurement]])) +
      geom_point() +
    xlab("Length of Stay (Days)") +
    ylab(label_title) +
    ggtitle(str_glue("Average LOS vs. {label_title}"))
```

```
      )
}
```

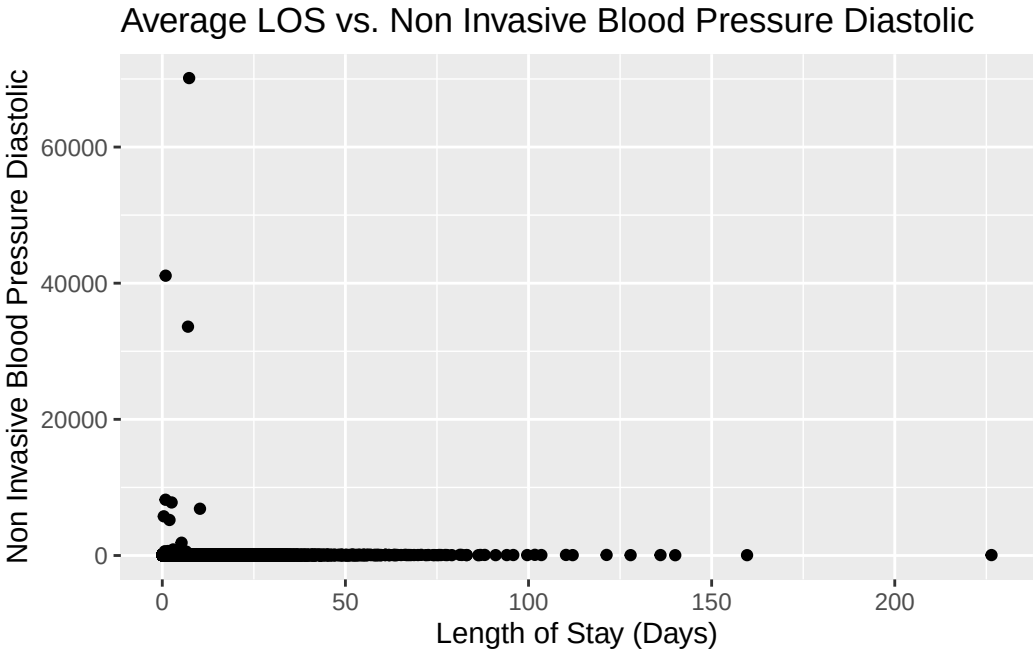
```
Warning: Removed 100 rows containing missing values or values outside the scale range
(`geom_point()`).
```



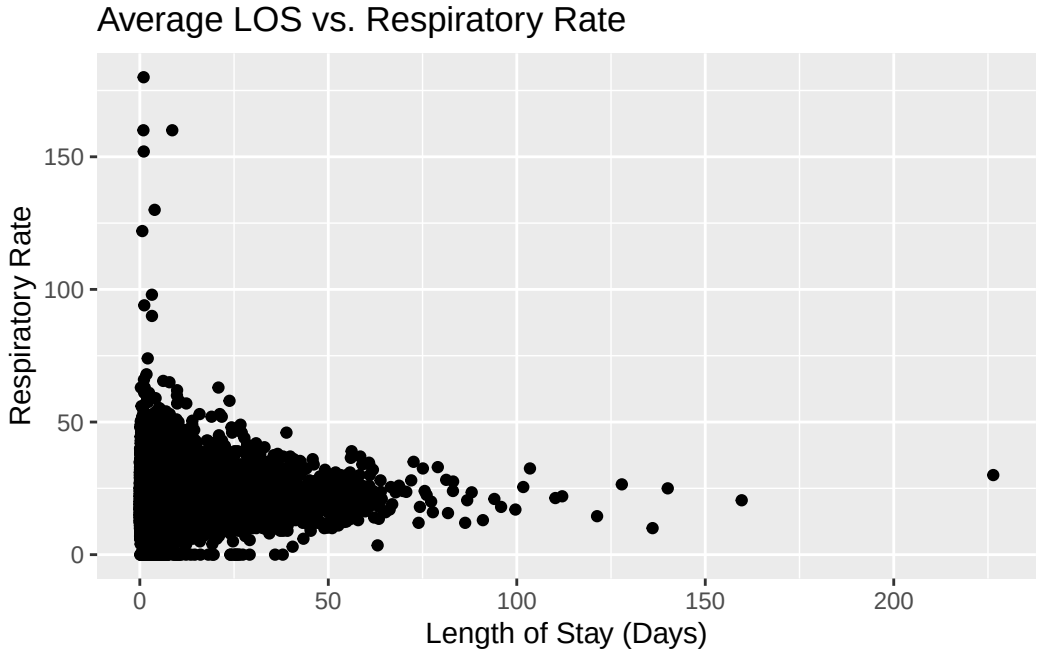

```
Warning: Removed 1384 rows containing missing values or values outside the scale range
(`geom_point()`).
```



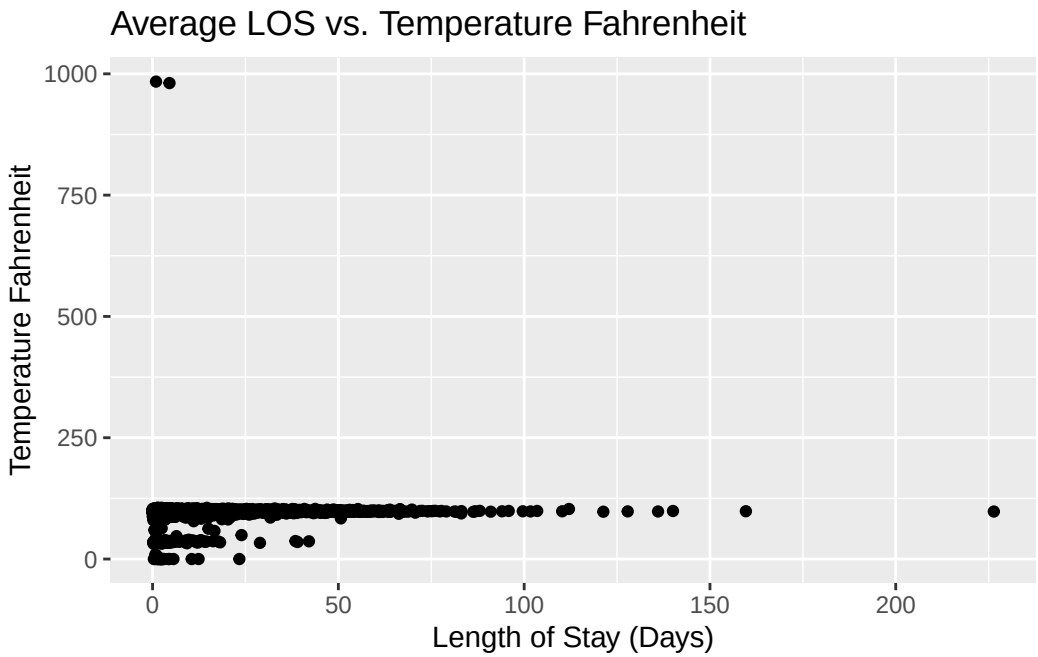

```
Warning: Removed 1389 rows containing missing values or values outside the scale range
(`geom_point()`).
```



```
Warning: Removed 212 rows containing missing values or values outside the scale range
(`geom_point()`).
```

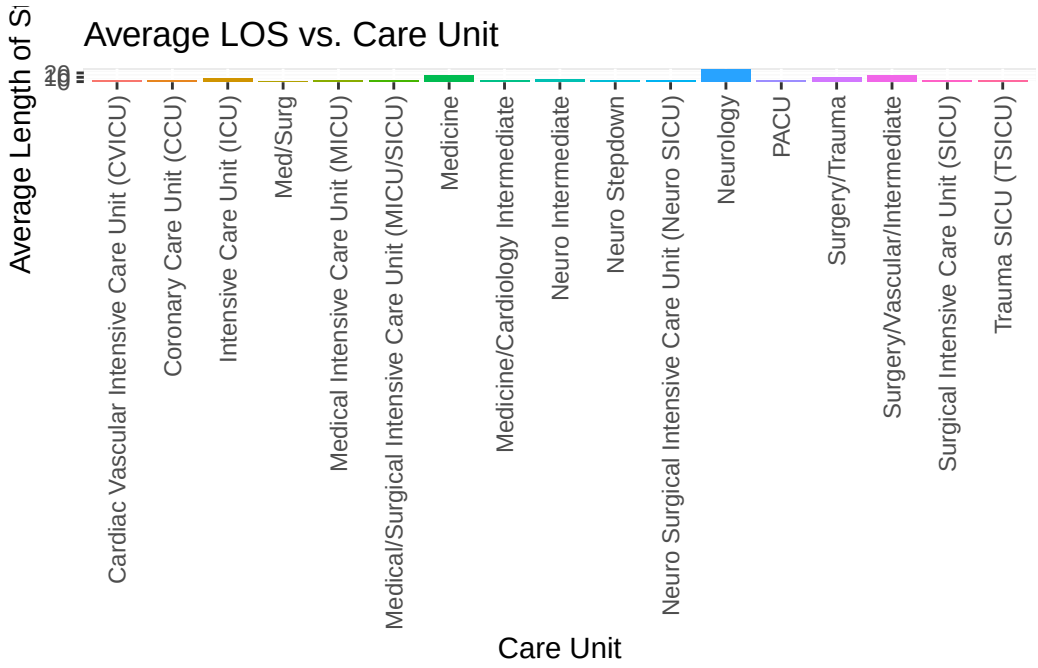



```
Warning: Removed 1690 rows containing missing values or values outside the scale range
(`geom_point()`).
```



These charts demonstrate strange outliers in the first measurements taken during patient ICU stays. For example, a heart rate exceeding 750 or blood pressure exceeding the 10000's should not be possible. These outliers could result from the chaos of the ICU and healthcare professionals mistakenly inputting in patient measurements.

```
#First ICU Unit
#We group by the first ICU Unit and take the mean of the length of stay.
los_vs_demo_first_ICU <- mimic_icu_cohort %>%
  group_by(first_careunit) %>%
  summarize(average_length_of_stay = mean(los, na.rm = TRUE)) %>%
  arrange(average_length_of_stay)

ggplot(los_vs_demo_first_ICU, mapping = aes(x = first_careunit,
  y = average_length_of_stay, fill = first_careunit)) +
  geom_bar(stat = "identity") +
  #Rotating the x-axis
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        legend.position="none") +
  ylab("Average Length of Stay (Days)") +
  xlab("Care Unit") +
  ggtitle("Average LOS vs. Care Unit")
```



This pattern demonstrates evident differences in average stay lengths depending on the care unit. This pattern results from different conditions requiring different lengths of treatment plans from their respective care units. For example, surgery care units require lengthier stay lengths for recovery times.