Biostat 203B Homework 4

Due Mar 9 @ 11:59PM

Kiana Mohamadinik and 205928003

Table of contents

Display machine information:

```r
sessionInfo()
```

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

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

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

time zone: America/Los_Angeles
```

```
tzcode source: internal

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

loaded via a namespace (and not attached):
 [1] compiler_4.3.0    fastmap_1.1.1     cli_3.6.3         tools_4.3.0
 [5] htmltools_0.5.8.1 rstudioapi_0.14   yaml_2.3.8        rmarkdown_2.29
 [9] knitr_1.45        jsonlite_1.8.8    xfun_0.50         digest_0.6.34
[13] rlang_1.1.4       evaluate_0.23
```

Display my machine memory.

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

```
Totalram:  16.000 GiB
Freeram:    1.007 GiB
```

Load database libraries and the tidyverse frontend:

```r
library(bigrquery)
```

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

```r
library(dbplyr)
library(DBI)
library(gt)
```

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

```r
library(gtsummary)
```

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

```r
library(tidyverse)
```

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

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

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

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

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.1
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::ident()  masks dbplyr::ident()
x dplyr::lag()    masks stats::lag()
x dplyr::sql()    masks dbplyr::sql()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

0.1 Q1. Compile the ICU cohort in HW3 from the Google BigQuery database

Below is an outline of steps. In this homework, we exclusively work with the BigQuery database and should not use any MIMIC data files stored on our local computer. Transform data as much as possible in BigQuery database and `collect()` the tibble **only at the end of Q1.7**.

0.1.1 Q1.1 Connect to BigQuery

Authenticate with BigQuery using the service account token. Please place the service account token (shared via BruinLearn) in the working directory (same folder as your qmd file). Do **not** ever add this token to your Git repository. If you do so, you will lose 50 points.

```
# path to the service account token
satoken <- "biostat-203b-2025-winter-4e58ec6e5579.json"
# BigQuery authentication using service account
bq_auth(path = satoken)
```

Connect to BigQuery database `mimiciv_3_1` in GCP (Google Cloud Platform), using the project billing account `biostat-203b-2025-winter`.

```
# connect to the BigQuery database `biostat-203b-2025-mimiciv_3_1`
con_bq <- dbConnect(
    bigrquery::bigquery(),
    project = "biostat-203b-2025-winter",
    dataset = "mimiciv_3_1",
    billing = "biostat-203b-2025-winter"
)
con_bq
```

```
<BigQueryConnection>
  Dataset: biostat-203b-2025-winter.mimiciv_3_1
  Billing: biostat-203b-2025-winter
```

List all tables in the `mimiciv_3_1` database.

```
dbListTables(con_bq)
```

```
 [1] "admissions"          "caregiver"           "chartevents"
 [4] "d_hcpcs"             "d_icd_diagnoses"     "d_icd_procedures"
 [7] "d_items"             "d_labitems"          "datetimeevents"
[10] "diagnoses_icd"       "drgcodes"            "emar"
[13] "emar_detail"         "hcpcsevents"         "icustays"
[16] "ingredientevents"    "inputevents"         "labevents"
[19] "microbiologyevents"  "omr"                 "outputevents"
[22] "patients"            "pharmacy"            "poe"
[25] "poe_detail"          "prescriptions"       "procedureevents"
[28] "procedures_icd"      "provider"            "services"
[31] "transfers"
```

0.1.2 Q1.2 `icustays#` data

Connect to the `icustays` table.

```
# full ICU stays table
icustays_tble <- tbl(con_bq, "icustays") |>
  arrange(subject_id, hadm_id, stay_id) |>
  # show_query() |>
  print(width = Inf)
```

```
# Source:     SQL [?? x 8]
# Database:   BigQueryConnection
# Ordered by: subject_id, hadm_id, stay_id
   subject_id  hadm_id  stay_id first_careunit
        <int>    <int>    <int> <chr>
 1   10000032 29079034 39553978 Medical Intensive Care Unit (MICU)
 2   10000690 25860671 37081114 Medical Intensive Care Unit (MICU)
 3   10000980 26913865 39765666 Medical Intensive Care Unit (MICU)
 4   10001217 24597018 37067082 Surgical Intensive Care Unit (SICU)
 5   10001217 27703517 34592300 Surgical Intensive Care Unit (SICU)
 6   10001725 25563031 31205490 Medical/Surgical Intensive Care Unit (MICU/SICU)
 7   10001843 26133978 39698942 Medical/Surgical Intensive Care Unit (MICU/SICU)
 8   10001884 26184834 37510196 Medical Intensive Care Unit (MICU)
 9   10002013 23581541 39060235 Cardiac Vascular Intensive Care Unit (CVICU)
10   10002114 27793700 34672098 Coronary Care Unit (CCU)
   last_careunit                                    intime
   <chr>                                            <dttm>
 1 Medical Intensive Care Unit (MICU)               2180-07-23 14:00:00
 2 Medical Intensive Care Unit (MICU)               2150-11-02 19:37:00
 3 Medical Intensive Care Unit (MICU)               2189-06-27 08:42:00
 4 Surgical Intensive Care Unit (SICU)              2157-11-20 19:18:02
 5 Surgical Intensive Care Unit (SICU)              2157-12-19 15:42:24
 6 Medical/Surgical Intensive Care Unit (MICU/SICU) 2110-04-11 15:52:22
 7 Medical/Surgical Intensive Care Unit (MICU/SICU) 2134-12-05 18:50:03
 8 Medical Intensive Care Unit (MICU)               2131-01-11 04:20:05
 9 Cardiac Vascular Intensive Care Unit (CVICU)     2160-05-18 10:00:53
10 Coronary Care Unit (CCU)                         2162-02-17 23:30:00
   outtime               los
   <dttm>              <dbl>
 1 2180-07-23 23:50:47 0.410
 2 2150-11-06 17:03:17 3.89
 3 2189-06-27 20:38:27 0.498
 4 2157-11-21 22:08:00 1.12
 5 2157-12-20 14:27:41 0.948
 6 2110-04-12 23:59:56 1.34
 7 2134-12-06 14:38:26 0.825
 8 2131-01-20 08:27:30 9.17
 9 2160-05-19 17:33:33 1.31
10 2162-02-20 21:16:27 2.91
# i more rows
```

0.1.3 Q1.3 `admissions` data

Connect to the `admissions` table.

```r
# # TODO
# admissions_tble <-
admissions_tble <- tbl(con_bq, "admissions") |>
    print(width = Inf)
```

```
# Source:   table<admissions> [?? x 16]
# Database: BigQueryConnection
   subject_id  hadm_id admittime           dischtime          
        <int>    <int> <dttm>              <dttm>             
 1   10106244 26713233 2147-05-09 10:34:00 2147-05-12 13:43:00
 2   13700703 20448599 2172-09-25 01:01:00 2172-10-03 13:25:00
 3   15443666 27961368 2168-12-30 23:30:00 2169-01-05 16:02:00
 4   16299919 26977065 2193-05-15 08:37:00 2193-05-17 16:03:00
 5   14149715 24191358 2181-10-25 19:37:00 2181-10-29 14:38:00
 6   14446098 20543394 2182-04-04 20:11:00 2182-05-07 19:00:00
 7   10584718 23485217 2165-02-12 15:41:00 2165-03-06 08:20:00
 8   12224488 25909420 2158-10-29 15:59:00 2158-11-01 15:45:00
 9   15845632 28189199 2124-10-05 02:44:00 2124-10-12 15:00:00
10   18131667 28337235 2195-11-18 02:58:00 2195-11-27 13:34:00
   deathtime           admission_type    admit_provider_id
   <dttm>              <chr>             <chr>            
 1 NA                  DIRECT EMER.      <NA>             
 2 NA                  OBSERVATION ADMIT <NA>             
 3 NA                  OBSERVATION ADMIT <NA>             
 4 NA                  OBSERVATION ADMIT <NA>             
 5 NA                  OBSERVATION ADMIT P00230           
 6 NA                  URGENT            P004G6           
 7 2165-03-06 08:20:00 EW EMER.          P004G6           
 8 NA                  EW EMER.          P004G6           
 9 NA                  EW EMER.          P004G6           
10 NA                  EW EMER.          P004G6           
   admission_location                     discharge_location       insurance
   <chr>                                  <chr>                    <chr>    
 1 PHYSICIAN REFERRAL                     HOME                     Private  
 2 EMERGENCY ROOM                         HOME                     Private  
 3 EMERGENCY ROOM                         HOME HEALTH CARE         Medicare 
 4 EMERGENCY ROOM                         HOSPICE                  Medicare 
 5 EMERGENCY ROOM                         SKILLED NURSING FACILITY Medicare 
 6 TRANSFER FROM HOSPITAL                  SKILLED NURSING FACILITY Medicare
 7 TRANSFER FROM SKILLED NURSING FACILITY DIED                     Medicare
 8 WALK-IN/SELF REFERRAL                  HOME                     Medicare
 9 PHYSICIAN REFERRAL                     HOME                     Private
10 PHYSICIAN REFERRAL                     HOME HEALTH CARE         Medicare
   language marital_status race                   edregtime
   <chr>    <chr>          <chr>                  <dttm>
 1 English  SINGLE         WHITE                  NA
 2 English  MARRIED        WHITE                  2172-09-24 17:38:00
 3 English  SINGLE         BLACK/AFRICAN AMERICAN 2168-12-30 11:19:00
 4 English  WIDOWED        BLACK/AFRICAN AMERICAN 2193-05-15 04:36:00
 5 English  SINGLE         WHITE                  2181-10-25 08:48:00
 6 English  MARRIED        WHITE                  NA
 7 English  MARRIED        WHITE                  NA
 8 English  SINGLE         WHITE - OTHER EUROPEAN 2158-10-28 20:22:00
 9 English  MARRIED        WHITE                  2124-10-04 19:30:00
10 English  SINGLE         WHITE                  2195-11-17 21:04:00
   edouttime           hospital_expire_flag
   <dttm>                             <int>
 1 NA                                     0
 2 2172-09-25 03:07:00                    0
 3 2168-12-31 01:22:00                    0
 4 2193-05-15 14:27:00                    0
 5 2181-10-26 15:18:00                    0
 6 NA                                     0
 7 NA                                     1
 8 2158-10-29 18:01:00                    0
 9 2124-10-05 04:10:00                    0
10 2195-11-18 04:51:00                    0
# i more rows
```

0.1.4 Q1.4 `patients` data

Connect to the `patients` table.

```r
# # TODO
# patients_tble <-
patients_tble <- tbl(con_bq, "patients") |>
    print(width = Inf)
```

```
# Source:   table<patients> [?? x 6]
# Database: BigQueryConnection
   subject_id gender anchor_age anchor_year anchor_year_group dod
        <int> <chr>       <int>       <int> <chr>             <date>
 1   10078138 F              18        2110 2017 - 2019       NA
 2   10180372 M              18        2110 2008 - 2010       NA
 3   10686175 M              18        2110 2011 - 2013       NA
 4   10851602 F              18        2110 2014 - 2016       NA
 5   10902424 F              18        2110 2017 - 2019       NA
 6   11092326 M              18        2110 2008 - 2010       NA
 7   11289691 F              18        2110 2017 - 2019       NA
 8   11595073 M              18        2110 2011 - 2013       NA
 9   11739764 F              18        2110 2017 - 2019       NA
10   11776346 F              18        2110 2008 - 2010       NA
# i more rows
```

0.1.5 Q1.5 `labevents` data

Connect to the `labevents` table and retrieve a subset that only contain subjects who appear in `icustays_tble` and the lab items listed in HW3. Only keep the last lab measurements (by `storetime`) before the ICU stay and pivot lab items to become variables/columns. Write all steps in *one* chain of pipes.

```r
# # TODO
# labevents_tble <-
target_lab_items <- c(
  50912,
  50971,
  50983,
  50902,
  50882,
  51221,
  51301,
  50931
)
labevents_tble <- tbl(con_bq, "labevents") |>
  filter(itemid %in% target_lab_items) |>
  arrange(subject_id, storetime, itemid)

labevents_tble <- labevents_tble |>
  inner_join(icustays_tble |> select(subject_id, stay_id, intime),
             by = "subject_id")
```

```r
labevents_tble <- labevents_tble |>
  filter(storetime < intime) |>
  mutate(valuenum = as.numeric(valuenum))

labevents_tble <- labevents_tble |>
  group_by(subject_id, stay_id, itemid) |>
  slice_max(order_by = storetime, n = 1, with_ties = FALSE) |>
  ungroup()

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

```
Warning: ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
```

```r
labevents_tble <- labevents_tble |>
  rename(
    creatinine = `50912`,
    potassium = `50971`,
    sodium = `50983`,
    chloride = `50902`,
    bicarbonate = `50882`,
    hematocrit = `51221`,
    wbc = `51301`,
    glucose = `50931`
  )

labevents_tble <- labevents_tble |>
  select(subject_id, stay_id, bicarbonate, chloride, creatinine,
         glucose, potassium, sodium, hematocrit, wbc) |>
  arrange(subject_id, stay_id)
labevents_tble <- labevents_tble |> collect()
```

```
Warning: ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
```

```
labevents_tble |> summarise(row_count = n())
```

```
# A tibble: 1 x 1
  row_count
      <int>
1     88086
```

```
labevents_tble
```

```
# A tibble: 88,086 x 10
   subject_id  stay_id bicarbonate chloride creatinine glucose potassium sodium
        <int>    <int>       <dbl>    <dbl>      <dbl>   <dbl>     <dbl>  <dbl>
 1   10000032 39553978          25       95        0.7     102       6.7    126
 2   10000690 37081114          26      100        1        85       4.8    137
 3   10000980 39765666          21      109        2.3      89       3.9    144
 4   10001217 34592300          30      104        0.5      87       4.1    142
 5   10001217 37067082          22      108        0.6     112       4.2    142
 6   10001725 31205490          NA       98       NA        NA       4.1    139
 7   10001843 39698942          28       97        1.3     131       3.9    138
 8   10001884 37510196          30       88        1.1     141       4.5    130
 9   10002013 39060235          24      102        0.9     288       3.5    137
10   10002114 34672098          18       NA        3.1      95       6.5    125
# i 88,076 more rows
# i 2 more variables: hematocrit <dbl>, wbc <dbl>
```

0.1.6 Q1.6 `chartevents` data

Connect to `chartevents` table and retrieve a subset that only contain subjects who appear in `icustays_tble` and the chart events listed in HW3. Only keep the first chart events (by `storetime`) during ICU stay and pivot chart events to become variables/columns. Write all steps in *one* chain of pipes. Similary to HW3, if a vital has multiple measurements at the first `storetime`, average them.

```
# # TODO
# chartevents_tble <-
vital_signs <- c(
  220045,
  220179,
  220180,
  223761,
  220210
)

chartevents_tble <- tbl(con_bq, "chartevents") |>
  filter(itemid %in% vital_signs) |>
  select(subject_id, stay_id, itemid, valuenum, storetime, charttime)

chartevents_tble <- chartevents_tble |>
  inner_join(
    icustays_tble |> select(subject_id, stay_id, intime, outtime),
    by = "stay_id"
  )

chartevents_tble <- chartevents_tble |>
  filter(storetime >= intime & storetime < outtime)

chartevents_tble <- chartevents_tble |>
  select(-subject_id_y) |>
  rename(subject_id = subject_id_x)

chartevents_tble <- chartevents_tble |>
  group_by(subject_id, stay_id, itemid) |>
  arrange(storetime) |>
  slice_min(order_by = storetime, n = 1, with_ties = TRUE) |>
  ungroup()

chartevents_tble <- chartevents_tble |>
  group_by(subject_id, stay_id, itemid) |>
  summarise(average_value = mean(valuenum, na.rm = TRUE), .groups = "drop")

chartevents_tble <- chartevents_tble |>
  pivot_wider(
    names_from = itemid,
    values_from = average_value,
    names_prefix = "vital_"
  )
```

```
Warning: ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
```

```
chartevents_tble <- chartevents_tble |>
  rename(
    heart_rate = vital_220045,
    non_invasive_blood_pressure_systolic = vital_220179,
    non_invasive_blood_pressure_diastolic = vital_220180,
    temperature_fahrenheit = vital_223761,
    respiratory_rate = vital_220210
  )

chartevents_tble <- chartevents_tble |>
  arrange(subject_id, stay_id)
chartevents_tble <- chartevents_tble |> collect()
```

```
Warning: ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
```

```
chartevents_tble |> summarise(row_count = n())
```

```
# A tibble: 1 x 1
  row_count
      <int>
1     94363
```

```
chartevents_tble
```

```
# A tibble: 94,363 x 7
   subject_id  stay_id non_invasive_blood_pressure_syst~1 temperature_fahrenheit
        <int>    <int>                              <dbl>                  <dbl>
 1   10000032 39553978                               84                     98.7
 2   10000690 37081114                              106                     97.7
 3   10000980 39765666                              154                     98  
 4   10001217 34592300                              156                     97.6
 5   10001217 37067082                              151                     98.5
 6   10001725 31205490                               73                     97.7
 7   10001843 39698942                              110                     97.9
 8   10001884 37510196                              174.                    98.1
 9   10002013 39060235                               98.5                   97.2
10   10002114 34672098                              112                     97.9
```

```
# i 94,353 more rows
# i abbreviated name: 1: non_invasive_blood_pressure_systolic
# i 3 more variables: respiratory_rate <dbl>,
#   non_invasive_blood_pressure_diastolic <dbl>, heart_rate <dbl>
```

0.1.7 Q1.7 Put things together

This step is similar to Q7 of HW3. Using *one* chain of pipes `|>` to perform following data wrangling steps: (i) start with the `icustays_tble`, (ii) merge in admissions and patients tables, (iii) keep adults only (age at ICU intime >= 18), (iv) merge in the labevents and chartevents tables, (v) `collect` the tibble, (vi) sort `subject_id`, `hadm_id`, `stay_id` and `print(width = Inf)`.

```r
# # TODO
# mimic_icu_cohort <-
icustays_age <- icustays_tble |>
  mutate(intime_year = lubridate::year(as.Date(intime)))

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

icustays_filtered <- icustays_tble |>
  left_join(age_at_intime, by = c("subject_id", "stay_id")) |>
  inner_join(
    patients_tble |> select(subject_id, gender, anchor_age, anchor_year,
                            anchor_year_group, dod),
    by = "subject_id"
  ) |>
  filter(age_at_intime >= 18)

icustays_filtered <- icustays_filtered |> collect()
```

```
Warning: ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
ORDER BY is ignored in subqueries without LIMIT
i Do you need to move arrange() later in the pipeline or use window_order() instead?
```

```r
admissions_selected <- admissions_tble |>
  select(subject_id, hadm_id, admittime, dischtime, deathtime,
         admission_type, admission_location, discharge_location,
         insurance, language, marital_status, edregtime, edouttime,
         hospital_expire_flag, admit_provider_id, race) |>
  collect()


mimic_icu_cohort <- icustays_filtered |>
  left_join(admissions_selected, by = c("subject_id", "hadm_id")) |>
  left_join(chartevents_tble, by = c("subject_id", "stay_id")) |>
  left_join(labevents_tble, by = c("subject_id", "stay_id")) |>
  distinct() |>
  arrange(subject_id, hadm_id, stay_id)

print(mimic_icu_cohort, width = Inf)
```

```
# A tibble: 94,458 x 41
   subject_id  hadm_id  stay_id first_careunit
        <int>    <int>    <int> <chr>
 1   10000032 29079034 39553978 Medical Intensive Care Unit (MICU)
 2   10000690 25860671 37081114 Medical Intensive Care Unit (MICU)
 3   10000980 26913865 39765666 Medical Intensive Care Unit (MICU)
 4   10001217 24597018 37067082 Surgical Intensive Care Unit (SICU)
 5   10001217 27703517 34592300 Surgical Intensive Care Unit (SICU)
 6   10001725 25563031 31205490 Medical/Surgical Intensive Care Unit (MICU/SICU)
 7   10001843 26133978 39698942 Medical/Surgical Intensive Care Unit (MICU/SICU)
 8   10001884 26184834 37510196 Medical Intensive Care Unit (MICU)
 9   10002013 23581541 39060235 Cardiac Vascular Intensive Care Unit (CVICU)
10   10002114 27793700 34672098 Coronary Care Unit (CCU)
   last_careunit                                    intime
   <chr>                                            <dttm>
 1 Medical Intensive Care Unit (MICU)               2180-07-23 14:00:00
 2 Medical Intensive Care Unit (MICU)               2150-11-02 19:37:00
 3 Medical Intensive Care Unit (MICU)               2189-06-27 08:42:00
 4 Surgical Intensive Care Unit (SICU)              2157-11-20 19:18:02
 5 Surgical Intensive Care Unit (SICU)              2157-12-19 15:42:24
 6 Medical/Surgical Intensive Care Unit (MICU/SICU) 2110-04-11 15:52:22
 7 Medical/Surgical Intensive Care Unit (MICU/SICU) 2134-12-05 18:50:03
 8 Medical Intensive Care Unit (MICU)               2131-01-11 04:20:05
 9 Cardiac Vascular Intensive Care Unit (CVICU)     2160-05-18 10:00:53
10 Coronary Care Unit (CCU)                         2162-02-17 23:30:00
```

```
   outtime               los age_at_intime gender anchor_age anchor_year
   <dttm>              <dbl>         <int> <chr>       <int>       <int>
 1 2180-07-23 23:50:47 0.410            52 F              52        2180
 2 2150-11-06 17:03:17 3.89             86 F              86        2150
 3 2189-06-27 20:38:27 0.498            76 F              73        2186
 4 2157-11-21 22:08:00 1.12             55 F              55        2157
 5 2157-12-20 14:27:41 0.948            55 F              55        2157
 6 2110-04-12 23:59:56 1.34             46 F              46        2110
 7 2134-12-06 14:38:26 0.825            76 M              73        2131
 8 2131-01-20 08:27:30 9.17             77 F              68        2122
 9 2160-05-19 17:33:33 1.31             57 F              53        2156
10 2162-02-20 21:16:27 2.91             56 M              56        2162
   anchor_year_group dod        admittime           dischtime
   <chr>             <date>     <dttm>              <dttm>
 1 2014 - 2016       2180-09-09 2180-07-23 12:35:00 2180-07-25 17:55:00
 2 2008 - 2010       2152-01-30 2150-11-02 18:02:00 2150-11-12 13:45:00
 3 2008 - 2010       2193-08-26 2189-06-27 07:38:00 2189-07-03 03:00:00
 4 2011 - 2013       NA         2157-11-18 22:56:00 2157-11-25 18:00:00
 5 2011 - 2013       NA         2157-12-18 16:58:00 2157-12-24 14:55:00
 6 2011 - 2013       NA         2110-04-11 15:08:00 2110-04-14 15:00:00
 7 2017 - 2019       2134-12-06 2134-12-05 00:10:00 2134-12-06 12:54:00
 8 2008 - 2010       2131-01-20 2131-01-07 20:39:00 2131-01-20 05:15:00
 9 2008 - 2010       NA         2160-05-18 07:45:00 2160-05-23 13:30:00
10 2020 - 2022       2162-12-11 2162-02-17 22:32:00 2162-03-04 15:16:00
   deathtime           admission_type              admission_location
   <dttm>              <chr>                       <chr>
 1 NA                  EW EMER.                    EMERGENCY ROOM
 2 NA                  EW EMER.                    EMERGENCY ROOM
 3 NA                  EW EMER.                    EMERGENCY ROOM
 4 NA                  EW EMER.                    EMERGENCY ROOM
 5 NA                  DIRECT EMER.                PHYSICIAN REFERRAL
 6 NA                  EW EMER.                    PACU
 7 2134-12-06 12:54:00 URGENT                      TRANSFER FROM HOSPITAL
 8 2131-01-20 05:15:00 OBSERVATION ADMIT           EMERGENCY ROOM
 9 NA                  SURGICAL SAME DAY ADMISSION PHYSICIAN REFERRAL
10 NA                  OBSERVATION ADMIT           PHYSICIAN REFERRAL
   discharge_location insurance language marital_status edregtime
   <chr>              <chr>     <chr>    <chr>          <dttm>
 1 HOME               Medicaid  English  WIDOWED        2180-07-23 05:54:00
 2 REHAB              Medicare  English  WIDOWED        2150-11-02 11:41:00
 3 HOME HEALTH CARE   Medicare  English  MARRIED        2189-06-27 06:25:00
 4 HOME HEALTH CARE   Private   Other    MARRIED        2157-11-18 17:38:00
 5 HOME HEALTH CARE   Private   Other    MARRIED        NA
 6 HOME               Private   English  MARRIED        NA
 7 DIED               Medicare  English  SINGLE         NA
 8 DIED               Medicare  English  MARRIED        2131-01-07 13:36:00
 9 HOME HEALTH CARE   Medicare  English  SINGLE         NA
10 HOME HEALTH CARE   Medicaid  English  <NA>           2162-02-17 19:35:00
   edouttime           hospital_expire_flag admit_provider_id
   <dttm>                             <int> <chr>
 1 2180-07-23 14:00:00                    0 P06OTX
 2 2150-11-02 19:37:00                    0 P26QQ4
 3 2189-06-27 08:42:00                    0 P06OTX
 4 2157-11-19 01:24:00                    0 P361ON
 5 NA                                     0 P276OU
 6 NA                                     0 P32W56
 7 NA                                     1 P67ATB
 8 2131-01-07 22:13:00                    1 P49AFC
 9 NA                                     0 P8286C
10 2162-02-17 23:30:00                    0 P46834
   race                   non_invasive_blood_pressure_systolic
   <chr>                                                 <dbl>
 1 WHITE                                                  84
 2 WHITE                                                 106
 3 BLACK/AFRICAN AMERICAN                                154
 4 WHITE                                                 151
 5 WHITE                                                 156
 6 WHITE                                                  73
 7 WHITE                                                 110
 8 BLACK/AFRICAN AMERICAN                                174.
 9 OTHER                                                  98.5
10 UNKNOWN                                               112
   temperature_fahrenheit respiratory_rate non_invasive_blood_pressure_diastolic
                    <dbl>            <dbl>                                 <dbl>
 1                   98.7             24                                    48
 2                   97.7             24.3                                  56.5
 3                   98               23.5                                 102
 4                   98.5             18                                    90
 5                   97.6             14                                    93.3
 6                   97.7             19                                    56
 7                   97.9             16.5                                  78
 8                   98.1             13                                    30.5
 9                   97.2             14                                    62
10                   97.9             21                                    80
   heart_rate bicarbonate chloride creatinine glucose potassium sodium
        <dbl>       <dbl>    <dbl>      <dbl>   <dbl>     <dbl>  <dbl>
 1       91            25      95       0.7     102       6.7    126
 2       78            26     100       1        85       4.8    137
 3       76            21     109       2.3      89       3.9    144
 4       86            22     108       0.6     112       4.2    142
 5       79.3          30     104       0.5      87       4.1    142
 6       86            NA      98      NA        NA       4.1    139
 7      124.           28      97       1.3     131       3.9    138
 8       49            30      88       1.1     141       4.5    130
 9       80            24     102       0.9     288       3.5    137
10      110.           18      NA       3.1      95       6.5    125
   hematocrit   wbc
        <dbl> <dbl>
 1       41.1   6.9
 2       36.1   7.1
 3       27.3   5.3
 4       38.1  15.7
 5       37.4   5.4
 6       NA    NA
 7       31.4  10.4
 8       39.7  12.2
 9       34.9   7.2
10       34.3  16.8
# i 94,448 more rows
```

```
glimpse(mimic_icu_cohort)
```

```
Rows: 94,458
Columns: 41
$ subject_id                            <int> 10000032, 10000690, 10000980, 10~
$ hadm_id                               <int> 29079034, 25860671, 26913865, 24~
$ stay_id                               <int> 39553978, 37081114, 39765666, 37~
$ first_careunit                        <chr> "Medical Intensive Care Unit (MI~
$ last_careunit                         <chr> "Medical Intensive Care Unit (MI~
$ intime                                <dttm> 2180-07-23 14:00:00, 2150-11-02~
$ outtime                               <dttm> 2180-07-23 23:50:47, 2150-11-06~
$ los                                   <dbl> 0.4102662, 3.8932523, 0.4975347,~
$ age_at_intime                         <int> 52, 86, 76, 55, 55, 46, 76, 77, ~
$ gender                                <chr> "F", "F", "F", "F", "F", "F", "M~
$ anchor_age                            <int> 52, 86, 73, 55, 55, 46, 73, 68, ~
$ anchor_year                           <int> 2180, 2150, 2186, 2157, 2157, 21~
$ anchor_year_group                     <chr> "2014 - 2016", "2008 - 2010", "2~
$ dod                                   <date> 2180-09-09, 2152-01-30, 2193-08~
$ admittime                                <dttm> 2180-07-23 12:35:00, 2150-11-02~
$ dischtime                                <dttm> 2180-07-25 17:55:00, 2150-11-12~
$ deathtime                                <dttm> NA, NA, NA, NA, NA, NA, 2134-12~
$ admission_type                           <chr> "EW EMER.", "EW EMER.", "EW EMER~
$ admission_location                       <chr> "EMERGENCY ROOM", "EMERGENCY ROO~
$ discharge_location                       <chr> "HOME", "REHAB", "HOME HEALTH CA~
$ insurance                                <chr> "Medicaid", "Medicare", "Medicar~
$ language                                 <chr> "English", "English", "English",~
$ marital_status                           <chr> "WIDOWED", "WIDOWED", "MARRIED",~
$ edregtime                                <dttm> 2180-07-23 05:54:00, 2150-11-02~
$ edouttime                                <dttm> 2180-07-23 14:00:00, 2150-11-02~
$ hospital_expire_flag                     <int> 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1,~
$ admit_provider_id                        <chr> "P060TX", "P26QQ4", "P060TX", "P~
$ race                                     <chr> "WHITE", "WHITE", "BLACK/AFRICAN~
$ non_invasive_blood_pressure_systolic  <dbl> 84.0, 106.0, 154.0, 151.0, 156.0~
$ temperature_fahrenheit                   <dbl> 98.7, 97.7, 98.0, 98.5, 97.6, 97~
$ respiratory_rate                         <dbl> 24.00000, 24.33333, 23.50000, 18~
$ non_invasive_blood_pressure_diastolic <dbl> 48.00000, 56.50000, 102.00000, 9~
$ heart_rate                               <dbl> 91.00000, 78.00000, 76.00000, 86~
$ bicarbonate                              <dbl> 25, 26, 21, 22, 30, NA, 28, 30, ~
$ chloride                                 <dbl> 95, 100, 109, 108, 104, 98, 97, ~
$ creatinine                               <dbl> 0.7, 1.0, 2.3, 0.6, 0.5, NA, 1.3~
$ glucose                                  <dbl> 102, 85, 89, 112, 87, NA, 131, 1~
$ potassium                                <dbl> 6.7, 4.8, 3.9, 4.2, 4.1, 4.1, 3.~
$ sodium                                   <dbl> 126, 137, 144, 142, 142, 139, 13~
$ hematocrit                               <dbl> 41.1, 36.1, 27.3, 38.1, 37.4, NA~
$ wbc                                      <dbl> 6.9, 7.1, 5.3, 15.7, 5.4, NA, 10~
```

0.1.8 Q1.8 Preprocessing

Perform the following preprocessing steps. (i) Lump infrequent levels into "Other" level for `first_careunit`, `last_careunit`, `admission_type`, `admission_location`, and `discharge_location`. (ii) Collapse the levels of `race` into `ASIAN`, `BLACK`, `HISPANIC`, `WHITE`, and `Other`. (iii) Create a new variable `los_long` that is `TRUE` when `los` is greater than or equal to 2 days. (iv) Summarize the data using `tbl_summary()`, stratified by `los_long`. Hint: `fct_lump_n` and `fct_collapse` from the `forcats` package are useful.

Hint: Below is a numerical summary of my tibble after preprocessing:

```r
library(forcats)
unique(mimic_icu_cohort$first_careunit)
```

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

```r
unique(mimic_icu_cohort$last_careunit)
```

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

```r
unique(mimic_icu_cohort$admission_type)
```

```
[1] "EW EMER."                    "DIRECT EMER."
```

```
[3] "URGENT"                      "OBSERVATION ADMIT"
[5] "SURGICAL SAME DAY ADMISSION" "ELECTIVE"
[7] "EU OBSERVATION"              "DIRECT OBSERVATION"
[9] "AMBULATORY OBSERVATION"
```

```r
unique(mimic_icu_cohort$admission_location)
```

```
 [1] "EMERGENCY ROOM"
 [2] "PHYSICIAN REFERRAL"
 [3] "PACU"
 [4] "TRANSFER FROM HOSPITAL"
 [5] "PROCEDURE SITE"
 [6] "TRANSFER FROM SKILLED NURSING FACILITY"
 [7] "WALK-IN/SELF REFERRAL"
 [8] "INFORMATION NOT AVAILABLE"
 [9] "CLINIC REFERRAL"
[10] "AMBULATORY SURGERY TRANSFER"
[11] "INTERNAL TRANSFER TO OR FROM PSYCH"
```

```r
unique(mimic_icu_cohort$discharge_location)
```

```
 [1] "HOME"                         "REHAB"
 [3] "HOME HEALTH CARE"             "DIED"
 [5] "CHRONIC/LONG TERM ACUTE CARE" "SKILLED NURSING FACILITY"
 [7] "PSYCH FACILITY"               "ACUTE HOSPITAL"
 [9] "OTHER FACILITY"               "HOSPICE"
[11] "AGAINST ADVICE"               NA
[13] "ASSISTED LIVING"              "HEALTHCARE FACILITY"
```

```r
unique(mimic_icu_cohort$race)
```

```
 [1] "WHITE"
 [2] "BLACK/AFRICAN AMERICAN"
 [3] "OTHER"
 [4] "UNKNOWN"
 [5] "UNABLE TO OBTAIN"
 [6] "WHITE - RUSSIAN"
 [7] "PORTUGUESE"
 [8] "BLACK/CAPE VERDEAN"
 [9] "HISPANIC/LATINO - SALVADORAN"
```

```
[10] "HISPANIC/LATINO - PUERTO RICAN"
[11] "ASIAN - SOUTH EAST ASIAN"
[12] "WHITE - OTHER EUROPEAN"
[13] "WHITE - BRAZILIAN"
[14] "HISPANIC OR LATINO"
[15] "BLACK/AFRICAN"
[16] "PATIENT DECLINED TO ANSWER"
[17] "HISPANIC/LATINO - GUATEMALAN"
[18] "ASIAN"
[19] "BLACK/CARIBBEAN ISLAND"
[20] "HISPANIC/LATINO - CUBAN"
[21] "ASIAN - CHINESE"
[22] "HISPANIC/LATINO - DOMINICAN"
[23] "ASIAN - KOREAN"
[24] "ASIAN - ASIAN INDIAN"
[25] "AMERICAN INDIAN/ALASKA NATIVE"
[26] "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER"
[27] "WHITE - EASTERN EUROPEAN"
[28] "HISPANIC/LATINO - CENTRAL AMERICAN"
[29] "HISPANIC/LATINO - HONDURAN"
[30] "HISPANIC/LATINO - COLUMBIAN"
[31] "SOUTH AMERICAN"
[32] "HISPANIC/LATINO - MEXICAN"
[33] "MULTIPLE RACE/ETHNICITY"
```

```r
summary(mimic_icu_cohort$los)
```

```
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
  0.00125   1.09621   1.96565   3.63002   3.86258 226.40308        14 
```

```r
library(forcats)
library(gtsummary)

mimic_icu_cohort_preprocessed <- mimic_icu_cohort |>
  mutate(
    first_careunit = fct_lump_n(first_careunit, n = 4, other_level = "Other"),
    last_careunit = fct_lump_n(last_careunit, n = 4, other_level = "Other"),
    admission_type = fct_lump_n(admission_type, n = 4, other_level = "Other"),
    admission_location = fct_lump_n(admission_location,
                                    n = 3, other_level = "Other"),
    discharge_location = fct_lump_n(discharge_location,
                                    n = 4, other_level = "Other")
  ) |>
  mutate(
    race = fct_collapse(
      race,
      ASIAN = c("ASIAN", "ASIAN - VIETNAMESE", "ASIAN - CHINESE",
                "ASIAN - FILIPINO", "ASIAN - OTHER", "ASIAN - SOUTH EAST ASIAN",
                "ASIAN - KOREAN", "ASIAN - ASIAN INDIAN"),
      BLACK = c("BLACK/AFRICAN AMERICAN", "BLACK/CAPE VERDEAN", "BLACK/HAITIAN",
                "BLACK/AFRICAN", "BLACK/CARIBBEAN ISLAND"),
      HISPANIC = c("HISPANIC OR LATINO", "HISPANIC/LATINO - PUERTO RICAN",
                   "HISPANIC/LATINO - DOMINICAN", "HISPANIC/LATINO - CUBAN",
                   "HISPANIC/LATINO - CENTRAL AMERICAN",
                   "HISPANIC/LATINO - SOUTH AMERICAN",
                   "HISPANIC/LATINO - MEXICAN",
                   "HISPANIC/LATINO - SALVADORAN",
                   "HISPANIC/LATINO - GUATEMALAN",
                   "HISPANIC/LATINO - HONDURAN", "HISPANIC/LATINO - COLUMBIAN"),
      WHITE = c("WHITE", "WHITE - RUSSIAN", "WHITE - BRAZILIAN",
                "WHITE - OTHER EUROPEAN", "WHITE - EASTERN EUROPEAN"),
      Other = c("AMERICAN INDIAN/ALASKA NATIVE",
                "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER",
                "MULTIPLE RACE/ETHNICITY", "UNABLE TO OBTAIN", "UNKNOWN",
                "PATIENT DECLINED TO ANSWER", "SOUTH AMERICAN", "OTHER",
                "PORTUGUESE")
    )
  ) |>
  mutate(
    los_long = los >= 2
  ) |>
  mutate(
    temperature_fahrenheit = ifelse(is.na(temperature_fahrenheit), NA,
                                    temperature_fahrenheit)
  )
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `race = fct_collapse(...)`.
Caused by warning:
! Unknown levels in `f`: ASIAN - VIETNAMESE, ASIAN - FILIPINO, ASIAN - OTHER, BLACK/HAITIAN,
```

```r
summary_table <- mimic_icu_cohort_preprocessed |>
  select(
    los_long, los, gender, race, age_at_intime, insurance,
    first_careunit, last_careunit, admission_type,
    admission_location, discharge_location, language,
    marital_status, hospital_expire_flag, dod,
    bicarbonate, chloride, creatinine, glucose, potassium, sodium, hematocrit,
    wbc, heart_rate, non_invasive_blood_pressure_systolic,
    non_invasive_blood_pressure_diastolic,
    temperature_fahrenheit, respiratory_rate
  ) |>
  tbl_summary(
    by = los_long,
    missing = "ifany"
  )
```

```
14 missing rows in the "los_long" column have been removed.
The following errors were returned during `tbl_summary()`:
x For variable `dod` (`los_long = FALSE`) and "p75" statistic: * not defined
  for "Date" objects
```

```r
summary_table
```

0.1.9 Q1.9 Save the final tibble

Save the final tibble to an R data file `mimic_icu_cohort.rds` in the `mimiciv_shiny` folder.

```r
# make a directory mimiciv_shiny
if (!dir.exists("mimiciv_shiny")) {
  dir.create("mimiciv_shiny")
}
# save the final tibble
mimic_icu_cohort |>
  write_rds("mimiciv_shiny/mimic_icu_cohort.rds", compress = "gz")
```

Close database connection and clear workspace.

```r
if (exists("con_bq")) {
  dbDisconnect(con_bq)
}
rm(list = ls())
```

Characteristic	**TRUE** N = 46,337[1]	**F**
los	3.9 (2.7, 6.8)	
gender		
F	20,106 (43%)	
M	26,231 (57%)	
race		
Other	8,036 (17%)	
ASIAN	1,369 (3.0%)	
BLACK	4,933 (11%)	
HISPANIC	1,687 (3.6%)	
WHITE	30,312 (65%)	
age_at_intime	67 (56, 77)	
insurance		
Medicaid	6,768 (15%)	
Medicare	26,330 (58%)	
No charge	5 (<0.1%)	
Other	1,091 (2.4%)	
Private	11,515 (25%)	
Unknown	628	
first_careunit		
Cardiac Vascular Intensive Care Unit (CVICU)	7,353 (16%)	
Medical Intensive Care Unit (MICU)	9,837 (21%)	
Medical/Surgical Intensive Care Unit (MICU/SICU)	6,667 (14%)	
Surgical Intensive Care Unit (SICU)	6,434 (14%)	
Other	16,046 (35%)	
last_careunit		
Cardiac Vascular Intensive Care Unit (CVICU)	7,353 (16%)	
Medical Intensive Care Unit (MICU)	9,837 (21%)	
Medical/Surgical Intensive Care Unit (MICU/SICU)	6,667 (14%)	
Surgical Intensive Care Unit (SICU)	6,434 (14%)	
Other	16,046 (35%)	
admission_type		
EW EMER.	23,012 (50%)	
OBSERVATION ADMIT	7,393 (16%)	
SURGICAL SAME DAY ADMISSION	4,001 (8.6%)	
URGENT	8,691 (19%)	
Other	3,240 (7.0%)	
admission_location		
EMERGENCY ROOM	17,058 (37%)	
PHYSICIAN REFERRAL	11,013 (24%)	
TRANSFER FROM HOSPITAL	13,904 (30%)	
Other	4,362 (9.4%)	
discharge_location		
DIED	6,884 (15%)	
HOME	6,879 (15%)	
HOME HEALTH CARE	10,620 (23%)	
SKILLED NURSING FACILITY	8,785 (19%)	
Other	13,092 (28%)	
Unknown	77	
language		

Although it is not a good practice to add big data files to Git, for grading purpose, please add `mimic_icu_cohort.rds` to your Git repository.

0.2 Q2. Shiny app

Develop a Shiny app for exploring the ICU cohort data created in Q1. The app should reside in the `mimiciv_shiny` folder. The app should contain at least two tabs. One tab provides easy access to the graphical and numerical summaries of variables (demographics, lab measurements, vitals) in the ICU cohort, using the `mimic_icu_cohort.rds` you curated in Q1. The other tab allows user to choose a specific patient in the cohort and display the patient's ADT and ICU stay information as we did in Q1 of HW3, by dynamically retrieving the patient's ADT and ICU stay information from BigQuery database. Again, do **not** ever add the BigQuery token to your Git repository. If you do so, you will lose 50 points.