# Biostat 203B Homework 4

**Due Mar 9 @ 11:59PM**

Ningke Zhang 705834790

Display machine information:

```
sessionInfo()
```

```
R version 4.4.2 (2024-10-31)
Platform: aarch64-apple-darwin20
Running under: macOS Sequoia 15.3.1

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

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

time zone: America/Los_Angeles
tzcode source: internal

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

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

Display my machine memory.

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

```
Totalram:   16.000 GiB
Freeram:   617.344 MiB
```

Load database libraries and the tidyverse frontend:

```
library(bigrquery)
library(dbplyr)
library(DBI)
library(gt)
library(gtsummary)
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 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 beco
```

```
library(forcats)
```

## 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**.

### 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"
```

### 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
```

### Q1.3 `admissions` data

Connect to the `admissions` table.

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

```
# Source:     SQL [?? x 16]
# Database:   BigQueryConnection
# Ordered by: subject_id, hadm_id
   subject_id  hadm_id admittime           dischtime           deathtime
        <int>    <int> <dttm>              <dttm>              <dttm>
 1   10000032 22595853 2180-05-06 22:23:00 2180-05-07 17:15:00 NA
 2   10000032 22841357 2180-06-26 18:27:00 2180-06-27 18:49:00 NA
 3   10000032 25742920 2180-08-05 23:44:00 2180-08-07 17:50:00 NA
 4   10000032 29079034 2180-07-23 12:35:00 2180-07-25 17:55:00 NA
 5   10000068 25022803 2160-03-03 23:16:00 2160-03-04 06:26:00 NA
 6   10000084 23052089 2160-11-21 01:56:00 2160-11-25 14:52:00 NA
 7   10000084 29888819 2160-12-28 05:11:00 2160-12-28 16:07:00 NA
 8   10000108 27250926 2163-09-27 23:17:00 2163-09-28 09:04:00 NA
 9   10000117 22927623 2181-11-15 02:05:00 2181-11-15 14:52:00 NA
10   10000117 27988844 2183-09-18 18:10:00 2183-09-21 16:30:00 NA
   admission_type    admit_provider_id admission_location     discharge_location
   <chr>             <chr>             <chr>                  <chr>
 1 URGENT            P49AFC            TRANSFER FROM HOSPITAL HOME
 2 EW EMER.          P784FA            EMERGENCY ROOM         HOME
 3 EW EMER.          P19UTS            EMERGENCY ROOM         HOSPICE
 4 EW EMER.          P06OTX            EMERGENCY ROOM         HOME
 5 EU OBSERVATION    P39NWO            EMERGENCY ROOM         <NA>
```

```
 6 EW EMER.          P42H7G            WALK-IN/SELF REFERRAL  HOME HEALTH CARE
 7 EU OBSERVATION    P35NE4            PHYSICIAN REFERRAL     <NA>
 8 EU OBSERVATION    P40JML            EMERGENCY ROOM         <NA>
 9 EU OBSERVATION    P47EY8            EMERGENCY ROOM         <NA>
10 OBSERVATION ADMIT P13ACE            WALK-IN/SELF REFERRAL  HOME HEALTH CARE
   insurance language marital_status race  edregtime
   <chr>     <chr>    <chr>          <chr> <dttm>
 1 Medicaid  English  WIDOWED        WHITE 2180-05-06 19:17:00
 2 Medicaid  English  WIDOWED        WHITE 2180-06-26 15:54:00
 3 Medicaid  English  WIDOWED        WHITE 2180-08-05 20:58:00
 4 Medicaid  English  WIDOWED        WHITE 2180-07-23 05:54:00
 5 <NA>      English  SINGLE         WHITE 2160-03-03 21:55:00
 6 Medicare  English  MARRIED        WHITE 2160-11-20 20:36:00
 7 Medicare  English  MARRIED        WHITE 2160-12-27 18:32:00
 8 <NA>      English  SINGLE         WHITE 2163-09-27 16:18:00
 9 Medicaid  English  DIVORCED       WHITE 2181-11-14 21:51:00
10 Medicaid  English  DIVORCED       WHITE 2183-09-18 08:41:00
   edouttime           hospital_expire_flag
   <dttm>                             <int>
 1 2180-05-06 23:30:00                    0
 2 2180-06-26 21:31:00                    0
 3 2180-08-06 01:44:00                    0
 4 2180-07-23 14:00:00                    0
 5 2160-03-04 06:26:00                    0
 6 2160-11-21 03:20:00                    0
 7 2160-12-28 16:07:00                    0
 8 2163-09-28 09:04:00                    0
 9 2181-11-15 09:57:00                    0
10 2183-09-18 20:20:00                    0
# i more rows
```

### Q1.4 `patients` data

Connect to the `patients` table.

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

```
# Source:     SQL [?? x 6]
```

```
# Database:   BigQueryConnection
# Ordered by: subject_id
   subject_id gender anchor_age anchor_year anchor_year_group dod
        <int> <chr>       <int>       <int> <chr>              <date>
 1   10000032 F              52        2180 2014 - 2016       2180-09-09
 2   10000048 F              23        2126 2008 - 2010       NA
 3   10000058 F              33        2168 2020 - 2022       NA
 4   10000068 F              19        2160 2008 - 2010       NA
 5   10000084 M              72        2160 2017 - 2019       2161-02-13
 6   10000102 F              27        2136 2008 - 2010       NA
 7   10000108 M              25        2163 2014 - 2016       NA
 8   10000115 M              24        2154 2017 - 2019       NA
 9   10000117 F              48        2174 2008 - 2010       NA
10   10000161 M              60        2163 2020 - 2022       NA
# i more rows
```

### 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.

```
# # TODO
labevents_tble <- tbl(con_bq, "labevents") |>
  select(subject_id, itemid, storetime, valuenum) |>
  filter(itemid %in% c(50912, 50971,
                       50983, 50902,
                       50882, 51221,
                       51301, 50931)) |>
  inner_join(icustays_tble, by = "subject_id") |>
  filter(storetime < intime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_max(order_by = storetime) |>
  ungroup() |>
  select(-c(hadm_id,
            storetime,
            intime,
            outtime,
            first_careunit,
            last_careunit, los)) |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
```

```
  rename(creatinine = "50912",
         potassium = "50971",
         sodium = "50983",
         chloride = "50902",
         bicarbonate = "50882",
         hematocrit = "51221",
         wbc = "51301",
         glucose = "50931") |>
# show_query() |>
print(width = Inf)
```

```
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?

# Source:   SQL [?? x 10]
# Database: BigQueryConnection
   subject_id  stay_id hematocrit bicarbonate   wbc creatinine chloride sodium
        <int>    <int>      <dbl>       <dbl> <dbl>      <dbl>    <dbl>  <dbl>
 1   10015931 37093652       29.7          27   5.3        1.8       98    138
 2   10048001 31935699       29.5          22   4.7        0.8      112    143
 3   10091141 36754593       35.6          23  11.8        5.4       97    133
 4   10118141 35472373       40.2          NA  19.1        1.8       NA     NA
 5   10128878 37360044       29.6          20   4.6        0.7       94    131
 6   10167691 35177099       31.2          27   7.5        1.3      100    138
 7   10169726 35268649       30.7          24   5.4        2.6      103    142
 8   10230862 36419736       28.8          22  17.9        1.2       99    136
 9   10240862 37475124       31.6          23   6.1        0.7      107    136
10   10288867 32495146       30.8          24  10.8        1.2      100    135
   potassium glucose
       <dbl>   <dbl>
 1       5.3      96
 2       4.4     131
 3       4.3     165
 4      NA       208
 5       3.1     289
 6       4.1      95
 7       4.6      88
 8       4.1     109
 9       4.2     145
```

```
10       4.1     130
# i more rows
```

```
labevents_tble |> count()
```

```
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?
```

```
# Source:   SQL [?? x 1]
# Database: BigQueryConnection
      n
  <int>
1 88086
```

### 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.

```
# # TODO
chartevents_tble <- tbl(con_bq, "chartevents") |>
  select(subject_id, stay_id, itemid, storetime, valuenum) |>
  filter(itemid %in% c(220045, 220179,
                       220180, 223761,
                       220210)) |>
  inner_join(icustays_tble, by = c("subject_id", "stay_id")) |>
  filter(storetime >= intime & storetime <= outtime) |>
  group_by(subject_id, stay_id, itemid) |>
  slice_min(order_by = storetime) |>
  ungroup() |>
  select(-c(hadm_id,
            storetime,
            intime,
            outtime,
            first_careunit,
            last_careunit,
            los)) |>
  pivot_wider(names_from = itemid, values_from = valuenum) |>
  rename(heart_rate = "220045",
```

```
         non_invasive_blood_pressure_systolic = "220179",
         non_invasive_blood_pressure_diastolic = "220180",
         temperature_fahrenheit = "223761",
         respiratory_rate = "220210") |>
# show_query() |>
print(width = Inf)
```

```
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?

# Source:   SQL [?? x 7]
# Database: BigQueryConnection
   subject_id  stay_id respiratory_rate non_invasive_blood_pressure_diastolic
        <int>    <int>            <dbl>                                 <dbl>
 1   10037928 38989978               16                                    59
 2   10080865 34903848               27                                    61
 3   10181673 36096029               22                                    35
 4   10215503 37836877               26                                    66
 5   10235995 33546260               15                                    90
 6   10263905 32177679               22                                    67
 7   10368327 36525878               18                                    62
 8   10439110 34201916               21                                    64
 9   10549546 33692245               17                                    64
10   10569306 32068207               18                                    66
   heart_rate temperature_fahrenheit non_invasive_blood_pressure_systolic
        <dbl>                 <dbl>                                 <dbl>
 1         76                  95.5                                   117
 2        102                  99.5                                    98
 3         73                  93.4                                    90
 4         82                 101.                                    140
 5         76                  98.4                                   135
 6         80                  97.7                                   118
 7         75                  97.6                                   118
 8         81                  97.6                                   151
 9         84                  99.2                                   135
10         64                  97.1                                   135
# i more rows
```

```
chartevents_tble |> count()
```

```
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?

# Source:   SQL [?? x 1]
# Database: BigQueryConnection
      n
  <int>
1 94363
```

### 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)`.

```
# # TODO
mimic_icu_cohort <- icustays_tble |>
  left_join(admissions_tble, by = c("subject_id", "hadm_id")) |>
  left_join(patients_tble, by = "subject_id") |>
  mutate(intime_age = year(intime) - anchor_year + anchor_age) |>
  filter(intime_age >= 18) |>
  left_join(labevents_tble, by = c("subject_id", "stay_id")) |>
  left_join(chartevents_tble, by = c("subject_id", "stay_id")) |>
  collect() |>
  arrange(subject_id, hadm_id, stay_id) |>
  print(width = Inf)
```

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

```
# 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 admittime           dischtime
   <dttm>              <dbl> <dttm>              <dttm>
 1 2180-07-23 23:50:47 0.410 2180-07-23 12:35:00 2180-07-25 17:55:00
 2 2150-11-06 17:03:17 3.89  2150-11-02 18:02:00 2150-11-12 13:45:00
 3 2189-06-27 20:38:27 0.498 2189-06-27 07:38:00 2189-07-03 03:00:00
 4 2157-11-21 22:08:00 1.12  2157-11-18 22:56:00 2157-11-25 18:00:00
 5 2157-12-20 14:27:41 0.948 2157-12-18 16:58:00 2157-12-24 14:55:00
 6 2110-04-12 23:59:56 1.34  2110-04-11 15:08:00 2110-04-14 15:00:00
 7 2134-12-06 14:38:26 0.825 2134-12-05 00:10:00 2134-12-06 12:54:00
 8 2131-01-20 08:27:30 9.17  2131-01-07 20:39:00 2131-01-20 05:15:00
 9 2160-05-19 17:33:33 1.31  2160-05-18 07:45:00 2160-05-23 13:30:00
10 2162-02-20 21:16:27 2.91  2162-02-17 22:32:00 2162-03-04 15:16:00
   deathtime           admission_type              admit_provider_id
   <dttm>              <chr>                       <chr>
 1 NA                  EW EMER.                    P06OTX
 2 NA                  EW EMER.                    P26QQ4
 3 NA                  EW EMER.                    P06OTX
 4 NA                  EW EMER.                    P361ON
```

```
 5 NA                  DIRECT EMER.               P276OU
 6 NA                  EW EMER.                   P32W56
 7 2134-12-06 12:54:00 URGENT                     P67ATB
 8 2131-01-20 05:15:00 OBSERVATION ADMIT          P49AFC
 9 NA                  SURGICAL SAME DAY ADMISSION P8286C
10 NA                  OBSERVATION ADMIT          P46834
   admission_location     discharge_location insurance language marital_status
   <chr>                  <chr>              <chr>     <chr>    <chr>
 1 EMERGENCY ROOM         HOME               Medicaid  English  WIDOWED
 2 EMERGENCY ROOM         REHAB              Medicare  English  WIDOWED
 3 EMERGENCY ROOM         HOME HEALTH CARE   Medicare  English  MARRIED
 4 EMERGENCY ROOM         HOME HEALTH CARE   Private   Other    MARRIED
 5 PHYSICIAN REFERRAL     HOME HEALTH CARE   Private   Other    MARRIED
 6 PACU                   HOME               Private   English  MARRIED
 7 TRANSFER FROM HOSPITAL DIED               Medicare  English  SINGLE
 8 EMERGENCY ROOM         DIED               Medicare  English  MARRIED
 9 PHYSICIAN REFERRAL     HOME HEALTH CARE   Medicare  English  SINGLE
10 PHYSICIAN REFERRAL     HOME HEALTH CARE   Medicaid  English  <NA>
   race                   edregtime           edouttime
   <chr>                  <dttm>              <dttm>
 1 WHITE                  2180-07-23 05:54:00 2180-07-23 14:00:00
 2 WHITE                  2150-11-02 11:41:00 2150-11-02 19:37:00
 3 BLACK/AFRICAN AMERICAN 2189-06-27 06:25:00 2189-06-27 08:42:00
 4 WHITE                  2157-11-18 17:38:00 2157-11-19 01:24:00
 5 WHITE                  NA                  NA
 6 WHITE                  NA                  NA
 7 WHITE                  NA                  NA
 8 BLACK/AFRICAN AMERICAN 2131-01-07 13:36:00 2131-01-07 22:13:00
 9 OTHER                  NA                  NA
10 UNKNOWN                2162-02-17 19:35:00 2162-02-17 23:30:00
   hospital_expire_flag gender anchor_age anchor_year anchor_year_group
                  <int> <chr>       <int>       <int> <chr>
 1                    0 F              52        2180 2014 - 2016
 2                    0 F              86        2150 2008 - 2010
 3                    0 F              73        2186 2008 - 2010
 4                    0 F              55        2157 2011 - 2013
 5                    0 F              55        2157 2011 - 2013
 6                    0 F              46        2110 2011 - 2013
 7                    1 M              73        2131 2017 - 2019
 8                    1 F              68        2122 2008 - 2010
 9                    0 F              53        2156 2008 - 2010
10                    0 M              56        2162 2020 - 2022
   dod        intime_age hematocrit bicarbonate   wbc creatinine chloride sodium
```

```
   <date>          <int>      <dbl>       <dbl> <dbl>      <dbl>    <dbl>  <dbl>
 1 2180-09-09         52       41.1          25   6.9        0.7       95    126
 2 2152-01-30         86       36.1          26   7.1        1        100    137
 3 2193-08-26         76       27.3          21   5.3        2.3      109    144
 4 NA                 55       38.1          22  15.7        0.6      108    142
 5 NA                 55       37.4          30   5.4        0.5      104    142
 6 NA                 46       NA            NA  NA         NA         98    139
 7 2134-12-06         76       31.4          28  10.4        1.3       97    138
 8 2131-01-20         77       39.7          30  12.2        1.1       88    130
 9 NA                 57       34.9          24   7.2        0.9      102    137
10 2162-12-11         56       34.3          18  16.8        3.1       NA    125
   potassium glucose respiratory_rate non_invasive_blood_pressure_diastolic
       <dbl>   <dbl>            <dbl>                                 <dbl>
 1       6.7     102               24                                    48
 2       4.8      85               27                                    63
 3       3.9      89               24                                   127
 4       4.2     112               18                                    90
 5       4.1      87               17                                    97
 6       4.1      NA               19                                    56
 7       3.9     131               17                                    85
 8       4.5     141               16                                    49
 9       3.5     288               14                                    70
10       6.5      95               22                                    80
   heart_rate temperature_fahrenheit non_invasive_blood_pressure_systolic
        <dbl>                  <dbl>                                <dbl>
 1         91                   98.7                                   84
 2         80                   97.7                                  107
 3         77                   98                                    158
 4         86                   98.5                                  151
 5         96                   97.6                                  167
 6         86                   97.7                                   73
 7        131                   97.9                                  112
 8         60                   98.1                                  180
 9         80                   97.2                                  104
10        111                   97.9                                  112
# i 94,448 more rows
```

### 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:

```
# # TODO

mimic_icu_cohort <- 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 = 4,
                                         other_level = "Other"),
         discharge_location = fct_lump(discharge_location, n = 4,
                                         other_level = "Other"),
  race = as.character(race),
  race = case_when(
    str_detect(race, regex("ASIAN", ignore_case = TRUE)) ~ "ASIAN",
    str_detect(race, regex("BLACK|AFRICAN", ignore_case = TRUE)) ~ "BLACK",
    str_detect(race, regex("HISPANIC|LATINO", ignore_case = TRUE)) ~ "HISPANIC",
    str_detect(race, regex("WHITE", ignore_case = TRUE)) ~ "WHITE",
    TRUE ~ "Other"),
  los_long = los >= 2) |>
  print(width = Inf)
```

```
# A tibble: 94,458 x 42
   subject_id  hadm_id  stay_id first_careunit
        <int>    <int>    <int> <fct>
 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 Other
   last_careunit                                      intime
```

```
   <fct>                                              <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 Other                                              2162-02-17 23:30:00
   outtime               los admittime           dischtime
   <dttm>              <dbl> <dttm>              <dttm>
 1 2180-07-23 23:50:47 0.410 2180-07-23 12:35:00 2180-07-25 17:55:00
 2 2150-11-06 17:03:17 3.89  2150-11-02 18:02:00 2150-11-12 13:45:00
 3 2189-06-27 20:38:27 0.498 2189-06-27 07:38:00 2189-07-03 03:00:00
 4 2157-11-21 22:08:00 1.12  2157-11-18 22:56:00 2157-11-25 18:00:00
 5 2157-12-20 14:27:41 0.948 2157-12-18 16:58:00 2157-12-24 14:55:00
 6 2110-04-12 23:59:56 1.34  2110-04-11 15:08:00 2110-04-14 15:00:00
 7 2134-12-06 14:38:26 0.825 2134-12-05 00:10:00 2134-12-06 12:54:00
 8 2131-01-20 08:27:30 9.17  2131-01-07 20:39:00 2131-01-20 05:15:00
 9 2160-05-19 17:33:33 1.31  2160-05-18 07:45:00 2160-05-23 13:30:00
10 2162-02-20 21:16:27 2.91  2162-02-17 22:32:00 2162-03-04 15:16:00
   deathtime           admission_type              admit_provider_id
   <dttm>              <fct>                       <chr>
 1 NA                  EW EMER.                    P06OTX
 2 NA                  EW EMER.                    P26QQ4
 3 NA                  EW EMER.                    P06OTX
 4 NA                  EW EMER.                    P3610N
 5 NA                  Other                       P276OU
 6 NA                  EW EMER.                    P32W56
 7 2134-12-06 12:54:00 URGENT                      P67ATB
 8 2131-01-20 05:15:00 OBSERVATION ADMIT           P49AFC
 9 NA                  SURGICAL SAME DAY ADMISSION P8286C
10 NA                  OBSERVATION ADMIT           P46834
   admission_location    discharge_location insurance language marital_status
   <fct>                 <fct>              <chr>     <chr>    <chr>
 1 EMERGENCY ROOM        HOME               Medicaid  English  WIDOWED
 2 EMERGENCY ROOM        Other              Medicare  English  WIDOWED
 3 EMERGENCY ROOM        HOME HEALTH CARE   Medicare  English  MARRIED
 4 EMERGENCY ROOM        HOME HEALTH CARE   Private   Other    MARRIED
 5 PHYSICIAN REFERRAL    HOME HEALTH CARE   Private   Other    MARRIED
 6 Other                 HOME               Private   English  MARRIED
```

```
 7 TRANSFER FROM HOSPITAL DIED               Medicare  English  SINGLE
 8 EMERGENCY ROOM         DIED               Medicare  English  MARRIED
 9 PHYSICIAN REFERRAL     HOME HEALTH CARE   Medicare  English  SINGLE
10 PHYSICIAN REFERRAL     HOME HEALTH CARE   Medicaid  English  <NA>
   race  edregtime           edouttime           hospital_expire_flag gender
   <chr> <dttm>              <dttm>                             <int> <chr>
 1 WHITE 2180-07-23 05:54:00 2180-07-23 14:00:00                    0 F
 2 WHITE 2150-11-02 11:41:00 2150-11-02 19:37:00                    0 F
 3 BLACK 2189-06-27 06:25:00 2189-06-27 08:42:00                    0 F
 4 WHITE 2157-11-18 17:38:00 2157-11-19 01:24:00                    0 F
 5 WHITE NA                  NA                                     0 F
 6 WHITE NA                  NA                                     0 F
 7 WHITE NA                  NA                                     1 M
 8 BLACK 2131-01-07 13:36:00 2131-01-07 22:13:00                    1 F
 9 Other NA                  NA                                     0 F
10 Other 2162-02-17 19:35:00 2162-02-17 23:30:00                    0 M
   anchor_age anchor_year anchor_year_group dod        intime_age hematocrit
        <int>       <int> <chr>             <date>          <int>      <dbl>
 1         52        2180 2014 - 2016       2180-09-09         52       41.1
 2         86        2150 2008 - 2010       2152-01-30         86       36.1
 3         73        2186 2008 - 2010       2193-08-26         76       27.3
 4         55        2157 2011 - 2013       NA                 55       38.1
 5         55        2157 2011 - 2013       NA                 55       37.4
 6         46        2110 2011 - 2013       NA                 46       NA
 7         73        2131 2017 - 2019       2134-12-06         76       31.4
 8         68        2122 2008 - 2010       2131-01-20         77       39.7
 9         53        2156 2008 - 2010       NA                 57       34.9
10         56        2162 2020 - 2022       2162-12-11         56       34.3
   bicarbonate   wbc creatinine chloride sodium potassium glucose
         <dbl> <dbl>      <dbl>    <dbl>  <dbl>     <dbl>   <dbl>
 1          25   6.9        0.7       95    126       6.7     102
 2          26   7.1        1        100    137       4.8      85
 3          21   5.3        2.3      109    144       3.9      89
 4          22  15.7        0.6      108    142       4.2     112
 5          30   5.4        0.5      104    142       4.1      87
 6          NA  NA         NA         98    139       4.1      NA
 7          28  10.4        1.3       97    138       3.9     131
 8          30  12.2        1.1       88    130       4.5     141
 9          24   7.2        0.9      102    137       3.5     288
10          18  16.8        3.1       NA    125       6.5      95
   respiratory_rate non_invasive_blood_pressure_diastolic heart_rate
              <dbl>                                 <dbl>      <dbl>
 1               24                                    48         91
```

```
 2               27                                    63         80
 3               24                                   127         77
 4               18                                    90         86
 5               17                                    97         96
 6               19                                    56         86
 7               17                                    85        131
 8               16                                    49         60
 9               14                                    70         80
10               22                                    80        111
   temperature_fahrenheit non_invasive_blood_pressure_systolic los_long
                    <dbl>                                <dbl> <lgl>
 1                   98.7                                   84 FALSE
 2                   97.7                                  107 TRUE
 3                   98                                    158 FALSE
 4                   98.5                                  151 FALSE
 5                   97.6                                  167 FALSE
 6                   97.7                                   73 FALSE
 7                   97.9                                  112 FALSE
 8                   98.1                                  180 TRUE
 9                   97.2                                  104 FALSE
10                   97.9                                  112 TRUE
# i 94,448 more rows
```

```
mimic_icu_cohort |>
    select(-c(subject_id,
              hadm_id,
              stay_id,
              intime,
              outtime,
              admittime,
              dischtime,
              deathtime,
              admit_provider_id,
              edregtime,
              edouttime,
              anchor_age,
              anchor_year,
              anchor_year_group)
           ) |>
  tbl_summary(
    by = los_long,
    statistic = list(all_continuous() ~ "{mean} ({sd})",
```

```
                    all_categorical() ~ "{n} / {N} ({p}%)")
  )
```

```
14 missing rows in the "los_long" column have been removed.
```

### Q1.9 Save the final tibble

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

```
# 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.

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

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.

## 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.

| **Characteristic** | **TRUE** N = 46,337[1] | **FALS** |
|---|---|---|
| first_careunit | | |
| Cardiac Vascular Intensive Care Unit (CVICU) | 7,353 / 46,337 (16%) | 7,416 |
| Medical Intensive Care Unit (MICU) | 9,837 / 46,337 (21%) | 10,862 |
| Medical/Surgical Intensive Care Unit (MICU/SICU) | 6,667 / 46,337 (14%) | 8,780 |
| Surgical Intensive Care Unit (SICU) | 6,434 / 46,337 (14%) | 6,574 |
| Other | 16,046 / 46,337 (35%) | 14,475 |
| last_careunit | | |
| Cardiac Vascular Intensive Care Unit (CVICU) | 7,353 / 46,337 (16%) | 7,416 |
| Medical Intensive Care Unit (MICU) | 9,837 / 46,337 (21%) | 10,862 |
| Medical/Surgical Intensive Care Unit (MICU/SICU) | 6,667 / 46,337 (14%) | 8,780 |
| Surgical Intensive Care Unit (SICU) | 6,434 / 46,337 (14%) | 6,574 |
| Other | 16,046 / 46,337 (35%) | 14,475 |
| los | 6.2 (6.8) | |
| admission_type | | |
| EW EMER. | 23,012 / 46,337 (50%) | 25,337 |
| OBSERVATION ADMIT | 7,393 / 46,337 (16%) | 6,638 |
| SURGICAL SAME DAY ADMISSION | 4,001 / 46,337 (8.6%) | 5,543 |
| URGENT | 8,691 / 46,337 (19%) | 6,683 |
| Other | 3,240 / 46,337 (7.0%) | 3,906 |
| admission_location | | |
| EMERGENCY ROOM | 17,058 / 46,337 (37%) | 20,443 |
| PHYSICIAN REFERRAL | 11,013 / 46,337 (24%) | 12,684 |
| TRANSFER FROM HOSPITAL | 13,904 / 46,337 (30%) | 10,400 |
| WALK-IN/SELF REFERRAL | 2,169 / 46,337 (4.7%) | 2,308 |
| Other | 2,193 / 46,337 (4.7%) | 2,272 |
| discharge_location | | |
| DIED | 6,884 / 46,260 (15%) | 4,436 |
| HOME | 6,879 / 46,260 (15%) | 15,210 |
| HOME HEALTH CARE | 10,620 / 46,260 (23%) | 13,422 |
| SKILLED NURSING FACILITY | 8,785 / 46,260 (19%) | 7,489 |
| Other | 13,092 / 46,260 (28%) | 6,779 |
| Unknown | 77 | |
| insurance | | |
| Medicaid | 6,768 / 45,709 (15%) | 7,469 |
| Medicare | 26,330 / 45,709 (58%) | 25,485 |
| No charge | 5 / 45,709 (<0.1%) | 3 / 4 |
| Other | 1,091 / 45,709 (2.4%) | 1,237 |
| Private | 11,515 / 45,709 (25%) | 13,018 |
| Unknown | 628 | |
| language | | |
| American Sign Language | 29 / 46,127 (<0.1%) | 34 / |
| Amharic | 14 / 46,127 (<0.1%) | 9 / 4 |
| Arabic | 87 / 46,127 (0.2%) | 62 / |
| Armenian | 12 / 46,127 (<0.1%) | 13 / |
| Bengali | 22 / 46,127 (<0.1%) | 12 / |
| Chinese | 550 / 46,127 (1.2%) | 611 / |
| English | 41,563 / 46,127 (90%) | 43,483 |
| French | 18 / 46,127 (<0.1%) | 14 / |
| Haitian | 375 / 46,127 (0.8%) | 252 |