Biostat 203B Homework 4

Due Mar 9 @ 11:59PM

AUTHOR
Jiayi Guo 206537584

Display machine information:

```
sessionInfo()
```

```
R version 4.3.3 (2024-02-29)
Platform: x86_64-apple-darwin20 (64-bit)
Running under: macOS 15.3.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-
x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-
x86_64/Resources/lib/libRlapack.dylib;  LAPACK version 3.11.0

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] htmlwidgets_1.6.4 compiler_4.3.3    fastmap_1.2.0     cli_3.6.3
 [5] tools_4.3.3       htmltools_0.5.8.1 rstudioapi_0.17.0 yaml_2.3.10
 [9] rmarkdown_2.28    knitr_1.49        jsonlite_1.8.9    xfun_0.48
[13] digest_0.6.37     rlang_1.1.4       evaluate_1.0.3
```

Display my machine memory.

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

```
Totalram:  16.000 GiB
Freeram:    1.595 GiB
```

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 ──
✔ dplyr     1.1.4     ✔ readr     2.1.5
✔ forcats   1.0.0     ✔ stringr   1.5.1
✔ ggplot2   3.5.1     ✔ tibble    3.2.1
✔ lubridate 1.9.3     ✔ tidyr     1.3.1
✔ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::ident()  masks dbplyr::ident()
✖ dplyr::lag()    masks stats::lag()
✖ dplyr::sql()    masks dbplyr::sql()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to
become errors
```

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"
```

```
d_items_tble <- tbl(con_bq, "d_items")
admissions_tble <- tbl(con_bq, "admissions")
patients_tble <- tbl(con_bq, "patients")
d_labitems_tble <- tbl(con_bq, "d_labitems")
```

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.

```
admissions_tble <- tbl(con_bq, "admissions") |>
  arrange(subject_id, hadm_id) |>
  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.

```
patients_tble <- tbl(con_bq, "patients") |>
  arrange(subject_id) |>
  print()
```

```
# 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.

steps Get the labevents

```
labevents_tble <- tbl(con_bq, "labevents") |>
  semi_join(icustays_tble, by = "hadm_id") |>
  filter(itemid %in% c(50912, 50971, 50983, 50902,
                       50882, 51221, 51301, 50931)) |>
  left_join(d_labitems_tble, by = "itemid") |>
  select(-subject_id) |>
  left_join(icustays_tble, by = "hadm_id") |>
  select(subject_id, stay_id, intime, itemid, storetime, valuenum, label) |>
  filter(storetime < intime) |>
  # Slice() isn't supported in SQL so I use distince instead
  arrange(stay_id, itemid, desc(storetime)) |>
  distinct(stay_id, itemid, .keep_all = TRUE) |>
  select(subject_id, stay_id, valuenum, label) |>
  pivot_wider(names_from = label, values_from = valuenum) |>
  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?
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
# Ordered by: stay_id
   subject_id  stay_id Sodium Glucose Creatinine Bicarbonate `White Blood Cells`
        <int>    <int>  <dbl>   <dbl>      <dbl>       <dbl>                <dbl>
 1   16430835 30014404    137      86        0.8          24                 NA
 2   15721974 30045061    139      91        0.8          27                  4.9
 3   16912984 30158579    141     105        0.9          24                  5.1
 4   13249026 30168063    148     128        1.2          23                 12
 5   12155780 30228591    134      89        1.3          20                  9
 6   18380416 30249802    138      89        0.7          23                  7
 7   18344051 30267249    139      96        0.7          26                  8.9
 8   11595344 30287697    131     120        0.7          21                 11.1
 9   14639454 30324540    140      68        0.6          27                 14.4
10   19069363 30351705    134     139        0.6          21                 13
   Hematocrit Potassium Chloride
        <dbl>     <dbl>    <dbl>
 1       46.2       3.9      102
 2       36.2       3.8      102
 3       35.4       3.9      105
 4       28.7       3.5      112
 5       35.9       4.8      100
 6       36.4       4.1      104
 7       27.3       4.2      102
 8       24.3       4.2       93
 9       27.5       4         105
10       29.4       4.1      104
# ℹ more rows
```

Unresolved hadm_id has lots of NA value, should I do something to this? lack lots of valuenum, is that normal? ### 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.

```
chartevents_tble <- tbl(con_bq, "chartevents") |>
  # Filter subjects appearing in icustay
  semi_join(icustays_tble, by = "stay_id") |>
  filter(itemid %in% c(220045, 220179, 220180, 223761, 220210)) |>
  left_join(d_items_tble, by = "itemid") |>
  select(-subject_id) |>
  select(-hadm_id) |>
  # Get intime and stay_id
  left_join(icustays_tble, by = "stay_id") |>
  # Divide every icu stay and item
  group_by(subject_id, stay_id, itemid) |>
  # Keep first measurement during icu stay
  filter(storetime >= intime) |>
  # Keep the smallest storetime
```

```
  filter(storetime == min(storetime)) |>
  # If there is only one measurement, the mean value is itself
  mutate(avg_value = mean(valuenum)) |>
  arrange(storetime) |>
  distinct(itemid, .keep_all = TRUE) |>
  ungroup() |>
  select(c("subject_id", "stay_id", "label", "avg_value")) |>
  pivot_wider(names_from = label, values_from = avg_value) |>
  arrange(stay_id) |>
  print(width = Inf)
```

```
Warning: Missing values are always removed in SQL aggregation functions.
Use `na.rm = TRUE` to silence this warning
This warning is displayed once every 8 hours.

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?
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
# Ordered by: stay_id
   subject_id  stay_id `Non Invasive Blood Pressure diastolic`
        <int>    <int>                                  <dbl>
 1   12466550 30000153                                   74
 2   13180007 30000213                                   62.5
 3   18421337 30000484                                   75
 4   12207593 30000646                                   68.5
 5   15726459 30000831                                  103
 6   12980335 30001148                                   48
 7   12168737 30001336                                   52
 8   17371178 30001396                                  103
 9   16513856 30001446                                   56
10   17461994 30001471                                   75
   `Temperature Fahrenheit` `Non Invasive Blood Pressure systolic` `Heart Rate`
                      <dbl>                                 <dbl>        <dbl>
 1                     99.1                                   136          104
 2                     97.4                                   165           74
 3                     96                                     101          106
 4                     98.6                                   111          100
 5                     98.6                                   115          122
```

```
 6                   95.6                                  102         80
 7                   98.5                                  110         70.5
 8                   98.8                                  169         86
 9                   98.1                                   75         82
10                   98.1                                  154         93.5
   `Respiratory Rate`
               <dbl>
 1                18
 2                20.5
 3                22
 4                28
 5                30.5
 6                 9.5
 7                30
 8                19
 9                22
10                16.5
# ℹ more rows
```

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)`.

```
mimic_icu_cohort <- icustays_tble |>
  left_join(admissions_tble, by = c("subject_id", "hadm_id")) |>
  left_join(patients_tble, by = c("subject_id")) |>
  left_join(labevents_tble, by = c("subject_id", "stay_id")) |>
  left_join(chartevents_tble, by = c("subject_id", "stay_id")) |>
  # compute age at intime
  mutate(age_intime = anchor_age + year(intime) - anchor_year) %>%
  # keep only patients aged over 18 at intime
  filter(age_intime > 18) %>%
  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?
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: 94,352 × 41
   subject_id  hadm_id  stay_id first_careunit last_careunit intime
        <int>    <int>    <int> <chr>          <chr>         <dttm>
 1   10270110 20171261 35854639 PACU           PACU          2134-03-25 03:32:02
 2   10270110 20171261 36372959 PACU           PACU          2134-03-24 01:31:39
 3   10270644 20019675 35548343 PACU           PACU          2159-12-03 16:20:31
 4   10368426 21588639 39194905 PACU           PACU          2164-12-30 13:29:21
 5   10464753 28216499 32421516 PACU           PACU          2183-01-10 20:51:04
 6   10640410 25898987 34344828 PACU           PACU          2112-02-03 12:55:23
 7   10691194 24438843 37799251 PACU           PACU          2147-06-01 17:38:48
 8   10710188 21362776 34067486 PACU           PACU          2147-06-22 11:48:40
 9   10710188 21362776 36638120 PACU           PACU          2147-05-28 16:18:40
10   10826759 28468289 37075137 PACU           PACU          2121-05-19 18:07:00
   outtime                 los admittime           dischtime
   <dttm>                <dbl> <dttm>              <dttm>
 1 2134-03-25 14:20:42  0.450  2134-03-22 04:57:00 2134-04-26 14:17:00
 2 2134-03-25 03:31:52  1.08   2134-03-22 04:57:00 2134-04-26 14:17:00
 3 2159-12-08 17:28:42  5.05   2159-12-03 01:17:00 2159-12-28 17:30:00
 4 2164-12-30 14:00:38  0.0217 2164-12-26 15:39:00 2165-01-03 16:30:00
 5 2183-01-11 22:58:45  1.09   2182-12-27 19:24:00 2183-01-27 17:39:00
 6 2112-02-08 15:14:54  5.10   2112-02-03 12:54:00 2112-02-19 18:00:00
 7 2147-06-01 17:58:44  0.0138 2147-04-25 08:30:00 2147-06-11 15:22:00
 8 2147-06-23 11:35:59  0.991  2147-05-28 16:17:00 2147-06-23 14:21:00
 9 2147-06-22 11:48:30 24.8    2147-05-28 16:17:00 2147-06-23 14:21:00
10 2121-05-20 16:32:39  0.934  2121-05-19 17:00:00 2121-05-24 12:30:00
   deathtime           admission_type    admit_provider_id
   <dttm>              <chr>             <chr>
 1 NA                  EW EMER.          P44KDZ
 2 NA                  EW EMER.          P44KDZ
 3 NA                  EW EMER.          P68D28
 4 NA                  EW EMER.          P46834
 5 NA                  OBSERVATION ADMIT P411FD
 6 NA                  OBSERVATION ADMIT P55X3P
 7 NA                  ELECTIVE          P93BYT
 8 2147-06-23 14:21:00 EW EMER.          P502T3
 9 2147-06-23 14:21:00 EW EMER.          P502T3
10 NA                  EW EMER.          P20PIB
   admission_location                    discharge_location      insurance
   <chr>                                 <chr>                   <chr>
 1 TRANSFER FROM HOSPITAL                HOSPICE                 Medicaid
 2 TRANSFER FROM HOSPITAL                HOSPICE                 Medicaid
```

```
 3 PHYSICIAN REFERRAL                    SKILLED NURSING FACILITY Medicare
 4 WALK-IN/SELF REFERRAL                 SKILLED NURSING FACILITY Medicare
 5 TRANSFER FROM HOSPITAL                HOSPICE                  Medicare
 6 CLINIC REFERRAL                       HOME HEALTH CARE         Private
 7 PHYSICIAN REFERRAL                    SKILLED NURSING FACILITY Medicare
 8 TRANSFER FROM SKILLED NURSING FACILITY DIED                    Medicare
 9 TRANSFER FROM SKILLED NURSING FACILITY DIED                    Medicare
10 TRANSFER FROM HOSPITAL                REHAB                    Medicare
   language marital_status race                   edregtime
   <chr>    <chr>          <chr>                  <dttm>
 1 English  MARRIED        WHITE                  2134-03-22 01:01:00
 2 English  MARRIED        WHITE                  2134-03-22 01:01:00
 3 English  DIVORCED       WHITE                  2159-12-02 19:45:00
 4 English  WIDOWED        WHITE                  2164-12-26 08:22:00
 5 English  MARRIED        UNABLE TO OBTAIN       2182-12-27 18:59:00
 6 English  MARRIED        BLACK/AFRICAN          2112-02-03 08:05:00
 7 English  WIDOWED        WHITE                  NA
 8 English  MARRIED        WHITE - OTHER EUROPEAN 2147-05-28 11:58:00
 9 English  MARRIED        WHITE - OTHER EUROPEAN 2147-05-28 11:58:00
10 English  SINGLE         WHITE - BRAZILIAN      2121-05-19 08:03:00
   edouttime           hospital_expire_flag gender anchor_age anchor_year
   <dttm>                             <int> <chr>       <int>       <int>
 1 2134-03-22 07:40:00                    0 M              78        2134
 2 2134-03-22 07:40:00                    0 M              78        2134
 3 2159-12-03 02:51:00                    0 F              84        2152
 4 2164-12-26 21:43:00                    0 M              80        2154
 5 2182-12-27 21:24:00                    0 M              86        2182
 6 2112-02-03 14:15:00                    0 F              44        2112
 7 NA                                     0 F              74        2144
 8 2147-05-28 18:23:00                    1 M              86        2147
 9 2147-05-28 18:23:00                    1 M              86        2147
10 2121-05-19 18:07:00                    0 F              77        2121
   anchor_year_group dod        Sodium Glucose Creatinine Bicarbonate
   <chr>             <date>      <dbl>   <dbl>      <dbl>       <dbl>
 1 2020 - 2022       2134-04-30    136     178        1            23
 2 2020 - 2022       2134-04-30    137      98        0.8          24
 3 2014 - 2016       2160-06-25    145      75        0.5          20
 4 2011 - 2013       2165-03-18    138     131        0.8          24
 5 2020 - 2022       2183-01-28    140     111        1.2          22
 6 2017 - 2019       NA             NA      NA       NA            NA
 7 2017 - 2019       2147-09-16    135      95        5            24
 8 2020 - 2022       2147-06-23    144     173        0.6          31
 9 2020 - 2022       2147-06-23     NA      NA       NA            NA
10 2020 - 2022       NA             NA      NA       NA            NA
   `White Blood Cells` Hematocrit Potassium Chloride
                 <dbl>      <dbl>     <dbl>    <dbl>
 1                13.3       19.6       3.8      105
 2                42.1       24.8       4        104
 3                 9.2       31.4       3.8      108
 4                 4.8       29.4       4        109
 5                12.8       31         3.6      107
```

```
 6               NA         NA        NA         NA
 7                8.7       26.2       4         94
 8               10.8       29.6       5        107
 9               NA         NA        NA         NA
10               NA         NA        NA         NA
   `Non Invasive Blood Pressure diastolic` `Temperature Fahrenheit`
                                     <dbl>                    <dbl>
 1                                      NA                     97.7
 2                                      NA                     96.7
 3                                      62                     97.9
 4                                      51                     NA
 5                                      63                     97.3
 6                                     107                     97.8
 7                                      86                     NA
 8                                      61                    101.
 9                                      61                     98.3
10                                      58                     98.5
   `Non Invasive Blood Pressure systolic` `Heart Rate` `Respiratory Rate`
                                    <dbl>        <dbl>               <dbl>
 1                                     NA           86                14
 2                                     NA          101                15
 3                                     92           55                16
 4                                    102           64                13
 5                                    106           96                20
 6                                    173           97                22
 7                                    127           98                10
 8                                     89           92                23
 9                                    104           62                23
10                                    116           77                16.5
   age_intime
        <int>
 1         78
 2         78
 3         91
 4         90
 5         87
 6         44
 7         77
 8         86
 9         86
10         77
# ℹ 94,342 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. **step(i)**

```
table(mimic_icu_cohort$first_careunit)
```

```
    Cardiac Vascular Intensive Care Unit (CVICU)
                                           14769
                         Coronary Care Unit (CCU)
                                           10772
                       Intensive Care Unit (ICU)
                                              33
                                         Med/Surg
                                               1
                Medical Intensive Care Unit (MICU)
                                           20672
Medical/Surgical Intensive Care Unit (MICU/SICU)
                                           15435
                                        Medicine
                                              16
                  Medicine/Cardiology Intermediate
                                               1
                               Neuro Intermediate
                                            5770
                                  Neuro Stepdown
                                            1420
 Neuro Surgical Intensive Care Unit (Neuro SICU)
                                            1749
                                       Neurology
                                               1
                                            PACU
                                             121
                                  Surgery/Trauma
                                              10
                    Surgery/Vascular/Intermediate
                                             145
               Surgical Intensive Care Unit (SICU)
                                           12993
                              Trauma SICU (TSICU)
                                           10444
```

```
mimic_icu_cohort$first_careunit <- fct_lump(mimic_icu_cohort$first_careunit,
                                            n = 7)
table(mimic_icu_cohort$last_careunit)
```

```
    Cardiac Vascular Intensive Care Unit (CVICU)
                                           14769
                         Coronary Care Unit (CCU)
                                           10772
                       Intensive Care Unit (ICU)
                                              33
```

```
                                        Med/Surg
                                               1
                    Medical Intensive Care Unit (MICU)
                                           20672
Medical/Surgical Intensive Care Unit (MICU/SICU)
                                           15435
                                        Medicine
                                              16
                     Medicine/Cardiology Intermediate
                                               1
                                 Neuro Intermediate
                                            5770
                                    Neuro Stepdown
                                            1420
 Neuro Surgical Intensive Care Unit (Neuro SICU)
                                            1749
                                       Neurology
                                               1
                                            PACU
                                             121
                                    Surgery/Trauma
                                              10
                       Surgery/Vascular/Intermediate
                                             145
                    Surgical Intensive Care Unit (SICU)
                                           12993
                                Trauma SICU (TSICU)
                                           10444
```

```
mimic_icu_cohort$last_careunit <- fct_lump(mimic_icu_cohort$last_careunit,
                                           n = 7)
table(mimic_icu_cohort$admission_type)
```

```
    AMBULATORY OBSERVATION                DIRECT EMER.
                        25                        3315
        DIRECT OBSERVATION                    ELECTIVE
                       237                        3027
           EU OBSERVATION                    EW EMER.
                       539                       48273
       OBSERVATION ADMIT SURGICAL SAME DAY ADMISSION
                     14020                        9540
                   URGENT
                     15376
```

```
mimic_icu_cohort$admission_type <- fct_lump(mimic_icu_cohort$admission_type,
                                           n = 4)
table(mimic_icu_cohort$admission_location)
```

```
           AMBULATORY SURGERY TRANSFER                       CLINIC REFERRAL
                                    76                                 1186
                        EMERGENCY ROOM               INFORMATION NOT AVAILABLE
                                 37443                                  229
    INTERNAL TRANSFER TO OR FROM PSYCH                                 PACU
                                    28                                  403
                   PHYSICIAN REFERRAL                        PROCEDURE SITE
                                 23677                                 1025
               TRANSFER FROM HOSPITAL TRANSFER FROM SKILLED NURSING FACILITY
                                 24298                                 1517
                WALK-IN/SELF REFERRAL
                                  4470
```

```
mimic_icu_cohort$admission_location <- fct_lump(
  mimic_icu_cohort$admission_location,
  n = 4)
table(mimic_icu_cohort$discharge_location)
```

```
             ACUTE HOSPITAL                AGAINST ADVICE
                        899                           840
           ASSISTED LIVING CHRONIC/LONG TERM ACUTE CARE
                         95                          6182
                       DIED           HEALTHCARE FACILITY
                      11325                            17
                       HOME              HOME HEALTH CARE
                      22030                         24036
                    HOSPICE                OTHER FACILITY
                       2546                           356
             PSYCH FACILITY                       REHAB
                        898                          8009
   SKILLED NURSING FACILITY
                      16273
```

```
mimic_icu_cohort$discharge_location <- fct_lump(
  mimic_icu_cohort$discharge_location,
  n = 4)
```

step(ii)

```
table(mimic_icu_cohort$race)
```

```
           AMERICAN INDIAN/ALASKA NATIVE
                                     198
                                   ASIAN
                                    1095
                  ASIAN - ASIAN INDIAN
                                     248
```

ASIAN - CHINESE	1060
ASIAN - KOREAN	73
ASIAN - SOUTH EAST ASIAN	407
BLACK/AFRICAN	431
BLACK/AFRICAN AMERICAN	8666
BLACK/CAPE VERDEAN	655
BLACK/CARIBBEAN ISLAND	621
HISPANIC OR LATINO	780
HISPANIC/LATINO - CENTRAL AMERICAN	73
HISPANIC/LATINO - COLUMBIAN	102
HISPANIC/LATINO - CUBAN	100
HISPANIC/LATINO - DOMINICAN	743
HISPANIC/LATINO - GUATEMALAN	227
HISPANIC/LATINO - HONDURAN	88
HISPANIC/LATINO - MEXICAN	87
HISPANIC/LATINO - PUERTO RICAN	1214
HISPANIC/LATINO - SALVADORAN	174
MULTIPLE RACE/ETHNICITY	74
NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER	131
OTHER	3130
PATIENT DECLINED TO ANSWER	514
PORTUGUESE	425
SOUTH AMERICAN	104
UNABLE TO OBTAIN	1874
UNKNOWN	8437
WHITE	

```
                                58840
                    WHITE - BRAZILIAN
                                  221
                WHITE - EASTERN EUROPEAN
                                  272
                  WHITE - OTHER EUROPEAN
                                 2308
                       WHITE - RUSSIAN
                                  980
```

```
mimic_icu_cohort <- mimic_icu_cohort |>
  mutate(
    race_total = case_when(
    str_detect(race, regex("ASIAN", ignore_case = TRUE)) ~ "ASIAN",
    str_detect(race, regex("BLACK", ignore_case = TRUE)) ~ "BLACK",
    str_detect(race, regex("HISPANIC", ignore_case = TRUE)) ~ "HISPANIC",
    str_detect(race, regex("WHITE", ignore_case = TRUE)) ~ "WHITE",
    TRUE ~ "Other"
    )
  ) |>
  select(-race) |>
  rename(race = race_total)
table(mimic_icu_cohort$race)
```

```
   ASIAN    BLACK HISPANIC    Other    WHITE
    2883    10373     3588    14887    62621
```

(iii)

```
mimic_icu_cohort <- mimic_icu_cohort |>
  mutate(
    los_long = if_else(los >= 2, TRUE, FALSE)
  )
```

(iv)

```
tbl_summary(mimic_icu_cohort,
            by = los_long,
            include = c(
              "first_careunit",
              "last_careunit",
              "admission_type",
              "admission_location",
              "discharge_location",
              "race"
            ))
```

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

Characteristic	FALSE N = 48,035[1]	TRUE N = 46,303[1]
first_careunit		
Cardiac Vascular Intensive Care Unit (CVICU)	7,414 (15%)	7,353 (16%)
Coronary Care Unit (CCU)	5,336 (11%)	5,432 (12%)
Medical Intensive Care Unit (MICU)	10,838 (23%)	9,830 (21%)
Medical/Surgical Intensive Care Unit (MICU/SICU)	8,769 (18%)	6,664 (14%)
Neuro Intermediate	2,073 (4.3%)	3,697 (8.0%)
Surgical Intensive Care Unit (SICU)	6,563 (14%)	6,429 (14%)
Trauma SICU (TSICU)	5,512 (11%)	4,932 (11%)
Other	1,530 (3.2%)	1,966 (4.2%)
last_careunit		
Cardiac Vascular Intensive Care Unit (CVICU)	7,414 (15%)	7,353 (16%)
Coronary Care Unit (CCU)	5,336 (11%)	5,432 (12%)
Medical Intensive Care Unit (MICU)	10,838 (23%)	9,830 (21%)
Medical/Surgical Intensive Care Unit (MICU/SICU)	8,769 (18%)	6,664 (14%)
Neuro Intermediate	2,073 (4.3%)	3,697 (8.0%)
Surgical Intensive Care Unit (SICU)	6,563 (14%)	6,429 (14%)
Trauma SICU (TSICU)	5,512 (11%)	4,932 (11%)
Other	1,530 (3.2%)	1,966 (4.2%)
admission_type		
EW EMER.	25,281 (53%)	22,988 (50%)
OBSERVATION ADMIT	6,631 (14%)	7,388 (16%)

[1] n (%)

Characteristic	FALSE N = 48,035[1]	TRUE N = 46,303[1]
SURGICAL SAME DAY ADMISSION	5,541 (12%)	3,999 (8.6%)
URGENT	6,679 (14%)	8,688 (19%)
Other	3,903 (8.1%)	3,240 (7.0%)
admission_location		
EMERGENCY ROOM	20,401 (42%)	17,042 (37%)
PHYSICIAN REFERRAL	12,667 (26%)	11,008 (24%)
TRANSFER FROM HOSPITAL	10,391 (22%)	13,896 (30%)
WALK-IN/SELF REFERRAL	2,306 (4.8%)	2,164 (4.7%)
Other	2,270 (4.7%)	2,193 (4.7%)
discharge_location		
DIED	4,435 (9.4%)	6,883 (15%)
HOME	15,167 (32%)	6,860 (15%)
HOME HEALTH CARE	13,415 (28%)	10,617 (23%)
SKILLED NURSING FACILITY	7,488 (16%)	8,785 (19%)
Other	6,761 (14%)	13,081 (28%)
Unknown	769	77
race		
ASIAN	1,515 (3.2%)	1,367 (3.0%)
BLACK	5,443 (11%)	4,930 (11%)
HISPANIC	1,903 (4.0%)	1,685 (3.6%)
Other	6,857 (14%)	8,025 (17%)
WHITE	32,317 (67%)	30,296 (65%)

[1] n (%)

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

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.