Biostat 203B Homework 4

Due Mar 9 @ 11:59PM

Sakshi Oza, 606542442

Display machine information:

```
sessionInfo()
```

```
R version 4.4.1 (2024-06-14)
Platform: aarch64-apple-darwin20
Running under: macOS Sonoma 14.4

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  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_4.4.1    fastmap_1.2.0     cli_3.6.4         tools_4.4.1
 [5] htmltools_0.5.8.1 rstudioapi_0.17.1 yaml_2.3.10       rmarkdown_2.29
 [9] knitr_1.49        jsonlite_1.9.1    xfun_0.51         digest_0.6.37
[13] rlang_1.1.5       evaluate_1.0.3
```

Display my machine memory.

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

```
Totalram:   8.000 GiB
Freeram:   62.047 MiB
```

Install the required libraries

```
install.packages(c(
  "shiny", "dplyr", "ggplot2", "gtsummary", "DBI",
  "bigrquery"
))
```

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     1.0.4
-- 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
```

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)

print(icustays_tble, 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(admissions_tble, 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(patients_tble, 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.

```r
# Load required libraries
library(bigrquery)
library(dbplyr)
library(dplyr)
library(tidyr)
library(stringr)

# Load d_labitems table and filter itemid
dlabitems_tble <- tbl(con_bq, "d_labitems") %>%
  filter(itemid %in% c(
    50912, 50971, 50983, 50902, 50882, 51221, 51301, 50931
  ))

# Query labevents from BigQuery
labs_data <- tbl(con_bq, "labevents") %>%
  select(subject_id, itemid, storetime, valuenum) %>%
  # Use semi_join to filter itemid from dlabitems_tble (without pulling into R)
  semi_join(dlabitems_tble, by = "itemid") %>%
  # Join with icustays table
  left_join(
```

```
    tbl(con_bq, "icustays") %>%
      select(subject_id, stay_id, intime),
    by = "subject_id"
  ) %>%
  # Filter for records before ICU intime
  filter(storetime < intime) %>%
  # Group by subject_id, stay_id, itemid
  group_by(subject_id, stay_id, itemid) %>%
  # Take the first row in each group
  slice_max(storetime, n = 1) %>%
  select(-storetime, -intime) %>%
  ungroup() %>%
  # Pivot wider to make itemid columns
  pivot_wider(names_from = itemid, values_from = valuenum) %>%
  # Rename specific columns
  rename(
    creatinine = `50912`,
    potassium = `50971`,
    sodium = `50983`,
    chloride = `50902`,
    bicarbonate = `50882`,
    hematocrit = `51221`,
    wbc = `51301`,
    glucose = `50931`
  ) %>%
  # Arrange by subject_id and stay_id
  arrange(subject_id, stay_id)

# Display the final dataframe
print(labs_data, width = Inf)
```

```
# Source:     SQL [?? x 10]
# Database:   BigQueryConnection
# Ordered by: subject_id, stay_id
   subject_id  stay_id potassium hematocrit glucose chloride creatinine sodium
        <int>    <int>     <dbl>      <dbl>   <dbl>    <dbl>      <dbl>  <dbl>
 1   10000032 39553978       6.7       41.1     102       95        0.7    126
 2   10000690 37081114       4.8       36.1      85      100          1    137
 3   10000980 39765666       3.9       27.3      89      109        2.3    144
 4   10001217 34592300       4.1       37.4      87      104        0.5    142
 5   10001217 37067082       4.2       38.1     112      108        0.6    142
 6   10001725 31205490       4.1       NA        NA       98         NA    139
```

```
 7   10001843 39698942       3.9       31.4     131       97         1.3     138
 8   10001884 37510196       4.5       39.7     141       88         1.1     130
 9   10002013 39060235       3.5       34.9     288      102         0.9     137
10   10002114 34672098       6.5       34.3      95       NA         3.1     125
   bicarbonate   wbc
         <dbl> <dbl>
 1          25   6.9
 2          26   7.1
 3          21   5.3
 4          30   5.4
 5          22  15.7
 6          NA  NA
 7          28  10.4
 8          30  12.2
 9          24   7.2
10          18  16.8
# i more rows
```

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")

# List of vital item IDs
vitals_itemids <- c(
  220045, # Heart rate
  220179, # Systolic non-invasive blood pressure
  220180, # Diastolic non-invasive blood pressure
  223761, # Body temperature (Fahrenheit)
  220210 # Respiratory rate
)

# Collect the ICU stays data
icustays_data <- icustays_tble %>%
  select(subject_id, stay_id, intime, outtime)

# Collect the chartevents data for the relevant item IDs
```

```r
vitals_data_raw <- chartevents_tble %>%
  select(
    subject_id, stay_id, itemid, charttime, value,
    valuenum, storetime
  ) %>%
  filter(itemid %in% vitals_itemids)

# Filter and join the data
# (after collecting both datasets)
vitals_data <- vitals_data_raw %>%
  inner_join(icustays_data, by = c("subject_id", "stay_id")) %>%
  filter(storetime >= intime & storetime <= outtime) %>%
  filter(!is.na(valuenum)) %>%
  group_by(subject_id, stay_id, itemid) %>%
  slice_min(storetime, with_ties = TRUE) %>%
  summarise(mean_value = mean(valuenum, na.rm = TRUE)) %>%
  ungroup() %>%
  pivot_wider(names_from = itemid, values_from = mean_value) %>%
  rename(
    heart_rate = `220045`,
    non_invasive_bloodpressure_systolic = `220179`,
    non_invasive_bloodpressure_diastolic = `220180`,
    temperature_fahrenheit = `223761`,
    respiratory_rate = `220210`
  ) %>%
  mutate(
    heart_rate = round(heart_rate, 1),
    non_invasive_bloodpressure_systolic = round(
      non_invasive_bloodpressure_systolic, 1
    ),
    non_invasive_bloodpressure_diastolic = round(
      non_invasive_bloodpressure_diastolic, 1
    ),
    temperature_fahrenheit = round(temperature_fahrenheit, 1),
    respiratory_rate = round(respiratory_rate, 1)
  ) %>%
  relocate(
    subject_id, stay_id, heart_rate,
    non_invasive_bloodpressure_diastolic,
    non_invasive_bloodpressure_systolic,
    respiratory_rate, temperature_fahrenheit
  ) %>%
```

```
  arrange(subject_id, stay_id)
```

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

```
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?
```

```
print(vitals_data, width = Inf)
```

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

```
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 7]
# Database:   BigQueryConnection
# Ordered by: subject_id, stay_id
   subject_id  stay_id heart_rate non_invasive_bloodpressure_diastolic
        <int>    <int>      <dbl>                                <dbl>
 1   10000032 39553978       91                                   48
 2   10000690 37081114       78                                   56.5
 3   10000980 39765666       76                                  102
 4   10001217 34592300       79.3                                 93.3
 5   10001217 37067082       86                                   90
 6   10001725 31205490       86                                   56
 7   10001843 39698942      124.                                  78
 8   10001884 37510196       49                                   30.5
 9   10002013 39060235       80                                   62
10   10002114 34672098      110.                                  80
   non_invasive_bloodpressure_systolic respiratory_rate temperature_fahrenheit
                                 <dbl>            <dbl>                   <dbl>
 1                                  84             24                     98.7
 2                                 106             24.3                   97.7
 3                                 154             23.5                   98
 4                                 156             14                     97.6
 5                                 151             18                     98.5
 6                                  73             19                     97.7
 7                                 110             16.5                   97.9
```

```
 8                          174.              13                  98.1
 9                           98.5             14                  97.2
10                          112               21                  97.9
# 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 %>%
  # Merge in admissions and patients tables
  left_join(patients_tble, by = "subject_id", copy = TRUE) %>%
  left_join(admissions_tble, by = c("hadm_id", "subject_id"), copy = TRUE) %>%
  # Keep only adults (age at ICU intime >= 18)
  mutate(
    intime_year = year(intime),
    age_intime = anchor_age + (intime_year - anchor_year)
  ) %>%
  filter(age_intime >= 18) %>%
  left_join(vitals_data,
    by = c("subject_id", "stay_id"),
    copy = TRUE
  ) %>%
  left_join(labs_data, by = c("subject_id", "stay_id"), copy = TRUE) %>%
  # Sort by subject_id, hadm_id, stay_id
  arrange(subject_id, hadm_id, stay_id) %>%
  # Remove the intermediate intime_year column
  select(-intime_year) %>%
  # Collect the tibble
  collect() %>%

# Print the final dataframe
print(mimic_icu_cohort, width = Inf)
```

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

```
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?
```

```
Warning: `...` must be empty in `format.tbl()`
Caused by error in `format_tbl()`:
! `...` must be empty.
x Problematic argument:
* ..1 = mimic_icu_cohort
i Did you forget to name an argument?
```

```
# 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 gender anchor_age anchor_year anchor_year_group
   <dttm>             <dbl> <chr>       <int>       <int> <chr>
 1 2180-07-23 23:50:47 0.410 F              52        2180 2014 - 2016
 2 2150-11-06 17:03:17 3.89  F              86        2150 2008 - 2010
 3 2189-06-27 20:38:27 0.498 F              73        2186 2008 - 2010
 4 2157-11-21 22:08:00 1.12  F              55        2157 2011 - 2013
 5 2157-12-20 14:27:41 0.948 F              55        2157 2011 - 2013
 6 2110-04-12 23:59:56 1.34  F              46        2110 2011 - 2013
 7 2134-12-06 14:38:26 0.825 M              73        2131 2017 - 2019
 8 2131-01-20 08:27:30 9.17  F              68        2122 2008 - 2010
 9 2160-05-19 17:33:33 1.31  F              53        2156 2008 - 2010
10 2162-02-20 21:16:27 2.91  M              56        2162 2020 - 2022
   dod        admittime           dischtime           deathtime
   <date>     <dttm>              <dttm>              <dttm>
 1 2180-09-09 2180-07-23 12:35:00 2180-07-25 17:55:00 NA
 2 2152-01-30 2150-11-02 18:02:00 2150-11-12 13:45:00 NA
 3 2193-08-26 2189-06-27 07:38:00 2189-07-03 03:00:00 NA
 4 NA         2157-11-18 22:56:00 2157-11-25 18:00:00 NA
 5 NA         2157-12-18 16:58:00 2157-12-24 14:55:00 NA
 6 NA         2110-04-11 15:08:00 2110-04-14 15:00:00 NA
 7 2134-12-06 2134-12-05 00:10:00 2134-12-06 12:54:00 2134-12-06 12:54:00
 8 2131-01-20 2131-01-07 20:39:00 2131-01-20 05:15:00 2131-01-20 05:15:00
 9 NA         2160-05-18 07:45:00 2160-05-23 13:30:00 NA
10 2162-12-11 2162-02-17 22:32:00 2162-03-04 15:16:00 NA
   admission_type              admit_provider_id admission_location
   <chr>                       <chr>             <chr>
 1 EW EMER.                    P06OTX            EMERGENCY ROOM
 2 EW EMER.                    P26QQ4            EMERGENCY ROOM
 3 EW EMER.                    P06OTX            EMERGENCY ROOM
 4 EW EMER.                    P3610N            EMERGENCY ROOM
 5 DIRECT EMER.                P276OU            PHYSICIAN REFERRAL
 6 EW EMER.                    P32W56            PACU
 7 URGENT                      P67ATB            TRANSFER FROM HOSPITAL
 8 OBSERVATION ADMIT           P49AFC            EMERGENCY ROOM
 9 SURGICAL SAME DAY ADMISSION P8286C            PHYSICIAN REFERRAL
10 OBSERVATION ADMIT           P46834            PHYSICIAN REFERRAL
   discharge_location insurance language marital_status race
   <chr>              <chr>     <chr>    <chr>          <chr>
 1 HOME               Medicaid  English  WIDOWED        WHITE
 2 REHAB              Medicare  English  WIDOWED        WHITE
```

```
 3 HOME HEALTH CARE   Medicare  English  MARRIED        BLACK/AFRICAN AMERICAN
 4 HOME HEALTH CARE   Private   Other    MARRIED        WHITE
 5 HOME HEALTH CARE   Private   Other    MARRIED        WHITE
 6 HOME               Private   English  MARRIED        WHITE
 7 DIED               Medicare  English  SINGLE         WHITE
 8 DIED               Medicare  English  MARRIED        BLACK/AFRICAN AMERICAN
 9 HOME HEALTH CARE   Medicare  English  SINGLE         OTHER
10 HOME HEALTH CARE   Medicaid  English  <NA>           UNKNOWN
   edregtime           edouttime           hospital_expire_flag age_intime
   <dttm>              <dttm>                             <int>      <int>
 1 2180-07-23 05:54:00 2180-07-23 14:00:00                    0         52
 2 2150-11-02 11:41:00 2150-11-02 19:37:00                    0         86
 3 2189-06-27 06:25:00 2189-06-27 08:42:00                    0         76
 4 2157-11-18 17:38:00 2157-11-19 01:24:00                    0         55
 5 NA                  NA                                     0         55
 6 NA                  NA                                     0         46
 7 NA                  NA                                     1         76
 8 2131-01-07 13:36:00 2131-01-07 22:13:00                    1         77
 9 NA                  NA                                     0         57
10 2162-02-17 19:35:00 2162-02-17 23:30:00                    0         56
   heart_rate non_invasive_bloodpressure_diastolic
        <dbl>                                <dbl>
 1       91                                   48
 2       78                                   56.5
 3       76                                  102
 4       86                                   90
 5       79.3                                 93.3
 6       86                                   56
 7      124.                                  78
 8       49                                   30.5
 9       80                                   62
10      110.                                  80
   non_invasive_bloodpressure_systolic respiratory_rate temperature_fahrenheit
                                 <dbl>            <dbl>                  <dbl>
 1                                84               24                     98.7
 2                               106               24.3                   97.7
 3                               154               23.5                   98
 4                               151               18                     98.5
 5                               156               14                     97.6
 6                                73               19                     97.7
 7                               110               16.5                   97.9
 8                               174.              13                     98.1
 9                                98.5             14                     97.2
```

```
10                              112              21                    97.9
   potassium hematocrit glucose chloride creatinine sodium bicarbonate   wbc
       <dbl>      <dbl>   <dbl>    <dbl>      <dbl>  <dbl>       <dbl> <dbl>
 1       6.7       41.1     102       95        0.7    126          25   6.9
 2       4.8       36.1      85      100        1      137          26   7.1
 3       3.9       27.3      89      109        2.3    144          21   5.3
 4       4.2       38.1     112      108        0.6    142          22  15.7
 5       4.1       37.4      87      104        0.5    142          30   5.4
 6       4.1       NA        NA       98       NA      139          NA  NA
 7       3.9       31.4     131       97        1.3    138          28  10.4
 8       4.5       39.7     141       88        1.1    130          30  12.2
 9       3.5       34.9     288      102        0.9    137          24   7.2
10       6.5       34.3      95       NA        3.1    125          18  16.8
# 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:

```
diagnosis_data <- tbl(con_bq, "diagnoses_icd")
transfers_data <- tbl(con_bq, "transfers")
procedures_data <- tbl(con_bq, "procedures_icd")
labevents_data <- tbl(con_bq, "labevents")
admission_data <- tbl(con_bq, "admissions")
patient_data <- tbl(con_bq, "patients")
d_icd_procedure <- tbl(con_bq, "d_icd_procedures")
chartevents_dataset <- tbl(con_bq, "chartevents")
d_icd_diagnosis <- tbl(con_bq, "d_icd_diagnoses")
```

```
# Load necessary libraries
library(dplyr)
library(forcats)
library(janitor)
```

```
Attaching package: 'janitor'

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

    chisq.test, fisher.test
```

```r
library(gtsummary)

preprocessed_data <- mimic_icu_cohort %>%
  # Lump infrequent levels into "Other" for specific categorical variables
  mutate(
    discharge_location = discharge_location %>%
      as_factor() %>%
      fct_drop(only = ""),
    # Lump infrequent levels into "Other"
    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"
    ),

    # Collapse levels of `race` into predefined categories
    race = fct_collapse(race,
      ASIAN = c(
        "ASIAN - ASIAN INDIAN", "PACIFIC ISLANDER",
        "ASIAN - CHINESE", "ASIAN - KOREAN",
        "ASIAN - SOUTH EAST ASIAN"
```

```
    ),
    BLACK = c(
      "BLACK/AFRICAN", "BLACK/AFRICAN AMERICAN",
      "BLACK/CAPE VERDEAN",
      "BLACK/CARIBBEAN ISLAND"
    ),
    HISPANIC = c(
      "HISPANIC OR LATINO",
      "HISPANIC/LATINO - CENTRAL AMERICAN",
      "HISPANIC/LATINO - COLUMBIAN",
      "HISPANIC/LATINO - CUBAN",
      "HISPANIC/LATINO - DOMINICAN",
      "HISPANIC/LATINO - GUATEMALAN",
      "HISPANIC/LATINO - HONDURAN",
      "HISPANIC/LATINO - MEXICAN",
      "HISPANIC/LATINO - PUERTO RICAN",
      "HISPANIC/LATINO - SALVADORAN"
    ),
    WHITE = c(
      "WHITE - BRAZILIAN",
      "WHITE - EASTERN EUROPEAN",
      "WHITE - OTHER EUROPEAN",
      "WHITE - RUSSIAN"
    ),
    Other = c(
      "AMERICAN INDIAN",
      "PATIENT DECLINED TO ANSWER",
      "PORTUGUESE", "SOUTH AMERICAN",
      "UNABLE TO OBTAIN", "UNKNOWN",
      "OTHER",
      "AMERICAN INDIAN/ALASKA NATIVE",
      "MULTIPLE RACE/ETHNICITY",
      "NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER"
    ),
  ),
  marital_status = marital_status %>% na_if(""),
  insurance = insurance %>% na_if(""),
  language = language %>% na_if(""),

  # Create a new variable `los_long` based on length of stay (`los`)
  los_long = los >= 2 # assuming 'los' is already in days
) %>%
```

```
  # Remove unnecessary variables
  select(
    -subject_id, -stay_id, -hadm_id, -intime, -outtime, -admittime,
    -dischtime, -deathtime, -admit_provider_id, -edregtime,
    -edouttime, -anchor_year_group, -anchor_age,
    -anchor_year
  )
```

```
Warning: There was 1 warning in `mutate()`.
i In argument: `race = fct_collapse(...)`.
Caused by warning:
! Unknown levels in `f`: PACIFIC ISLANDER, AMERICAN INDIAN
```

```
# Summarize the data stratified by `los_long` using `tbl_summary`
summary_table <- tbl_summary(
  preprocessed_data,
  by = "los_long", # Stratify by `los_long`
  statistic = list(
    all_categorical() ~ "{n} ({p}%)",
    # Adjust statistics for categorical/continuous variables
    all_continuous() ~ "{median} ({p25}),{p75}"
  ),
  missing = "ifany" # Handle missing data
)
```

```
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
```

```
# Print summary
summary_table
```

Q1.9 Save the final tibble

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

Characteristic	**TRUE** N = 46,337[1]	**FA**
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%)	
los	3.9 (2.7),6.8	
gender		
F	20,106 (43%)	
M	26,231 (57%)	
dod	2155-09-06 (2135-07-16),2175-10-08	2155-1
Unknown	25,846	
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		
HOME	6,879 (15%)	
HOME HEALTH CARE	10,620 (23%)	
DIED	6,884 (15%)	
SKILLED NURSING FACILITY	8,785 (19%)	
Other	13,092 (28%)	
Unknown	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	
language		
American Sign Language	29 (<0.1%)	
Amharic	14 (<0.1%)	
Arabic	87 (0.2%)	
Armenian	12 (<0.1%)	
Bengali	22 (<0.1%)	

```
# 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.