# Assignment 2

## Recommended R Reading

[R for Data Science, Chapter 5 Data Transformation](https://r4ds.had.co.nz/transform.html)

* Sections “Filter rows with filter()”, “Arrange columns with arrange()”, “Select columns with select()”, “Grouped summaries with summarise()”

## Other R Resources

[The Tidyverse Cookbook](https://rstudio-education.github.io/tidyverse-cookbook/)

* Edited by one of the authors of *R for Data Science*, this online text provides example code for the common data processing tasks illustrated in class and much more

## Task 1: Practice using the pipe operator to chain data processing steps

Rewrite the following blocks of code using the Tidyverse pipe operator %>%



core1p\_adults <- filter(core1p, !is.na(AGE) & AGE >= 18)

core1p\_adults\_with\_nonelective <- mutate(

core1p\_adults, nonelective = ifelse(ATYPE != 3, 1, 0))



core1p\_ed <- filter(core1p, HCUP\_ED %in% c(2, 4))

core1p\_ed\_pr\_vars <- select(

core1p\_ed,

KEY, VisitLink, DaysToEvent, LOS, LOS\_X, starts\_with("I10\_PR")

)



core1p\_no\_missing\_los <- mutate(core1p, LOS = ifelse(is.na(LOS), LOS\_X, LOS))

core1p\_mcare\_mcaid <- filter(core1p\_no\_missing\_los, PAY1 %in% c(1, 2))

core1p\_mcare\_mcaid\_grouped <- group\_by(core1p\_mcare\_mcaid, PAY1)

mcare\_mcaid\_summary <- summarise(

core1p\_mcare\_mcaid\_grouped,

n\_admits = n(),

avg\_los = mean(LOS)

)

## Task 2: Practice using select, filter, group\_by, and summarise

Using the 1% sample of the Florida SID

1. produce a data set containing KEY, VisitLink, DaysToEvent, and all of the variables specifically named in the ‘Core File Variables’ slide in the first R lecture;
2. produce a data set containing all discharges between the first quarter of 2016 and the third quarter of 2019.
3. produce a data set summarizing the number of admissions by discharge status (using the variable DISPUNIFORM) (Hint: recall that we can define a variable with the count of admissions with n\_admits = n()).

## Task 3: Practice applying exclusions one-by-one

Replicate the exclusions of [Hammond *et al.* (2020)](https://doi.org/10.1161/STROKEAHA.120.029318) using the 1% sample of the Florida SID:

Of 851 429 admissions from 2012 to 2017, we excluded 39 564 missing data for rurality (n=2942), age (n=40), gender (n=133), zip level income (n=16,129), insurance status (n=1319), transfer status (n=4675), discharge destination (n=400), length of stay (n=41), or total charges (n=13 885). We also excluded people<age 18 (n=1937). (p 2132)

Specifically,

1. Exclude for missing rurality (assume they used the variable PL\_NCHS to define rurality).
2. Exclude for missing age or age < 18.
3. Exclude for missing gender.
4. Exclude for missing zip code-level income (assume they used the variable ZIPINC\_QRTL).
5. Exclude for missing primary payer (PAY1).
6. Exclude for missing transfer status (TRAN\_IN and TRAN\_OUT). Don’t worry about any other transfer-related exclusions.
7. Exclude for missing discharge destination (DISPUNIFORM).
8. Exclude for missing length of stay (use LOS\_X here).
9. Exclude for missing total charges (use TOTCHG\_X).

Apply the exclusions one-by-one, and report the number excluded at each step as a comment in your code. For example, to determine the number excluded at step 1, take the number of discharge records in the data set before applying the exclusion (e.g., nrow(core1p)) and subtract from it the number of discharges in the data set after applying the exclusion (e.g., nrow(core1p\_no\_missing\_rurality)). Although not as efficient as applying the exclusions in a single step, you’ll want to apply exclusions one-by-one for your project so you can record the number excluded for each cause.

## Task 4: Practice using group by processing to generate summary statistics

Find the proportion of hospitalizations among adult patients ending in death by rural/urban classification along with the number of hospitalizations in each rural/urban category. That is, using the 1% sample data set, compute the number of admissions and the proportion of admissions with DIED == 1 grouped by PL\_NCHS and first filtered by adult (nonmissing AGE and AGE >= 18).