BIOS 721 Data Management: Data Manipulation

This exercise is to help you practice various data manipulation tasks in R. We will look at merging, dealing with dates, and pivot tables.

Merging and Dates

Use the **pat_demo** and **pat_comorbid** datasets on Canvas to complete the following:

- Read in both datasets
- Merge the datasets together
- Find duplicates
- Remove duplicates
- Fix the DOB variable
- Remove the old DOB variable
- Print the top of the tibble

```
## # A tibble: 49 x 9
##    PATID birth_date race  ethnicity   financialclass hypertension CHF   diabetes
##    <chr> <date>     <chr> <chr>       <chr>          <chr>        <chr> <chr>
##  1 Z5278 1958-07-01 Other non-Hispan~ Private        N            N     N
##  2 Z9101 1965-10-21 White non-Hispan~ Private        Y            N     N
##  3 Z1649 1982-01-21 White non-Hispan~ Medicare       N            N     N
##  4 Z3946 1950-01-30 White non-Hispan~ Medicare       Y            N     Y
##  5 Z4582 1989-09-10 White non-Hispan~ Private        Y            N     N
##  6 Z1392 1957-08-15 White non-Hispan~ Private        N            N     N
##  7 Z8517 1977-07-17 White non-Hispan~ Medicare       Y            N     N
##  8 Z4318 1973-10-23 White non-Hispan~ Medicare       Y            N     N
##  9 Z3359 1951-08-27 White non-Hispan~ Private        Y            N     N
## 10 Z9306 1984-10-29 Other non-Hispan~ Medicare       N            N     Y
## # i 39 more rows
## # i 1 more variable: CKD <chr>
```

Pivot Tables

Start by reading in the *utilization.csv* dataset (on Canvas). This dataset contains information on the monthly utilization of the Duke Health system for select patients.

Data Dictionary
PATID: Patient identifier
Month: The month for which utilization measures are recorded
EDvisits: The number of ED visits for each patient
Admissions: Total number of hospital admissions
Readmissions: Total number of readmissions (an admission with 30 days of a previous admission)
ACSCadmissions: Total number of Ambulatory Care Sensitive Conditions Admissions (avoidable admissions for things like high blood pressure)
PCPvisit: Total number of visits to a primary care provider

After reading in the data, remove all the NAs from the data. Then convert your data from long to wide, using *only* the ED visit variable (drop all other utilization variables).

```
## # A tibble: 19 x 38
##    PATID `5/1/2019` `6/1/2019` `7/1/2019` `8/1/2019` `9/1/2019` `10/1/2019`
##    <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>       <dbl>
##  1 41216          0          0          0          0          0           0
##  2 41927          0          0          0          0          0           0
##  3 44126          0          0          0          0          0           0
##  4 46821          0          0          0          0          1           0
##  5 53086          0          0          0          0          0           0
##  6 53215         NA         NA         NA         NA         NA          NA
##  7 53524          0          0          2          1          0           0
##  8 53630          0          0          0          0          0           0
##  9 58644          0          0          0          0          0           0
## 10 68282          0          0          0          0          0           0
## 11 76970          0          1          2          2          2           2
## 12 79553          0          0          0          0          1           0
## 13 81665          0          0          0          0          0           0
## 14 81885          0          0          0          0          0           0
## 15 83450          0          0          0          0          0           0
## 16 86945          0          0          0          0          0           0
## 17 95158          0          0          0          0          0           0
## 18 95597          0          2          1          0          0           0
## 19 97371          0          0          0          0          0           0
## # i 31 more variables: `11/1/2019` <dbl>, `12/1/2019` <dbl>, `1/1/2020` <dbl>,
## #   `2/1/2020` <dbl>, `3/1/2020` <dbl>, `4/1/2020` <dbl>, `5/1/2020` <dbl>,
## #   `6/1/2020` <dbl>, `7/1/2020` <dbl>, `8/1/2020` <dbl>, `9/1/2020` <dbl>,
## #   `10/1/2020` <dbl>, `11/1/2020` <dbl>, `12/1/2020` <dbl>, `1/1/2021` <dbl>,
## #   `2/1/2021` <dbl>, `3/1/2021` <dbl>, `4/1/2021` <dbl>, `5/1/2021` <dbl>,
## #   `6/1/2021` <dbl>, `7/1/2021` <dbl>, `8/1/2021` <dbl>, `9/1/2021` <dbl>,
## #   `10/1/2021` <dbl>, `11/1/2021` <dbl>, `12/1/2021` <dbl>, ...
```

What do you notice about the wide data?

Now convert your wide data back to being long data.

```
## # A tibble: 703 x 3
##    PATID Month     EDvisits
##    <dbl> <chr>        <dbl>
##  1 41216 5/1/2019         0
##  2 41216 6/1/2019         0
##  3 41216 7/1/2019         0
##  4 41216 8/1/2019         0
##  5 41216 9/1/2019         0
##  6 41216 10/1/2019        0
##  7 41216 11/1/2019        0
##  8 41216 12/1/2019        0
##  9 41216 1/1/2020         1
## 10 41216 2/1/2020         0
## # i 693 more rows
```

Compare to original data

```
## # A tibble: 601 x 3
##    PATID Month     EDvisits
##    <dbl> <chr>        <dbl>
##  1 41216 5/1/2019         0
##  2 41216 6/1/2019         0
##  3 41216 7/1/2019         0
##  4 41216 8/1/2019         0
##  5 41216 9/1/2019         0
##  6 41216 10/1/2019        0
##  7 41216 11/1/2019        0
##  8 41216 12/1/2019        0
##  9 41216 1/1/2020         1
## 10 41216 2/1/2020         0
## # i 591 more rows
```

What issues do you notice about your new long data set?