

## Assignment 2: Data transformation

Due Sunday, October 10, 2021 @ 11:59pm

**Public Health 280.346, Fall 2021** Your assignment for this module is to explore the full National Medical Expenditures Survey (NMES) dataset using the functions in the `dplyr` package. You will also re-create the graph you reproduced in Module 1 to make use of what you've learned about data transformation. Your assignment has three parts:

- (1) Use the functions in the `dplyr` package to answer the following questions about our NMES data set:
  - How many people in the dataset have an MSCD?
  - What was the highest medical expenditure in this data?
  - What was the highest medical expenditure for someone without an MSCD?
  - How old is the youngest person with an MSCD?
  - How old is the oldest person without an MSCD?
- (2) Write R code to reproduce the following tables that compare people with and without an MSCD for the variables age, bmi, educate, poor, and female.

```
## # A tibble: 2 x 4
##   mscd      n mean_age sd_age
##   <chr> <int>   <dbl>  <dbl>
## 1 No    3801    45.7   18.1
## 2 Yes   277     67.3   12.8

## # A tibble: 2 x 4
##   mscd      n mean_bmi sd_bmi
##   <chr> <int>   <dbl>  <dbl>
## 1 No   3684    25.5   5.06
## 2 Yes   270    25.5   4.40

## `summarise()` has grouped output by 'mscd'. You can override using the `.groups` argument.

## # A tibble: 8 x 4
## # Groups:   mscd [2]
##   mscd educate      n percent
##   <chr> <fct>   <int>   <dbl>
## 1 No   CollGrad  650    17.1
## 2 No   SomeColl  753    19.8
## 3 No   HSGrad    1906   50.1
## 4 No   Other     492    12.9
## 5 Yes  CollGrad   30    10.8
## 6 Yes  SomeColl   39    14.1
## 7 Yes  HSGrad    148   53.4
## 8 Yes  Other     60    21.7

## `summarise()` has grouped output by 'mscd'. You can override using the `.groups` argument.

## # A tibble: 4 x 4
## # Groups:   mscd [2]
##   mscd poor      n percent
##   <chr> <fct>   <int>   <dbl>
## 1 No   Not Poor  2988   78.6
## 2 No   Poor     813   21.4
## 3 Yes  Not Poor  116   41.9
## 4 Yes  Poor     161   58.1

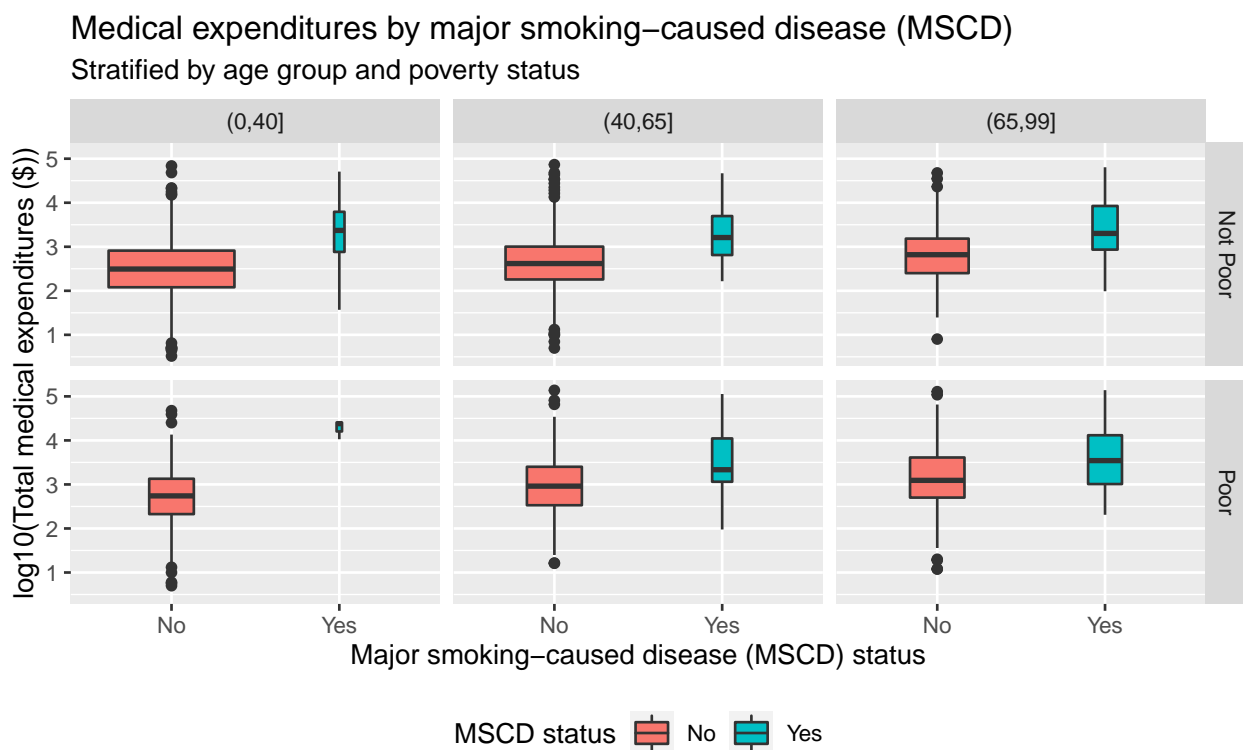
## `summarise()` has grouped output by 'mscd'. You can override using the `.groups` argument.
```

```
## # A tibble: 4 x 4
## # Groups:   mscd [2]
##   mscd female    n percent
##   <chr> <fct> <int>   <dbl>
## 1 No    Male    1446    38.0
## 2 No    Female  2355    62.0
## 3 Yes    Male     137    49.5
## 4 Yes    Female   140    50.5
```

- (3) Now that we've learned how to transform our data, we can improve our boxplot from Module 1 that shows the relationship between medical expenditure and MSCD status by working the log10-transformed expenditures rather than medical expenditures on the dollar scale. Reproduce the two plots below **exactly** using the `dplyr` and `ggplot2` packages in R. Which of these two plots makes it the easiest for you to compare individuals with MSCD to “otherwise similar” individuals without MSCD? Explain your choice.

**Hint:** For the second plot, you will need to create a new variable that combines the information from the `ageCat` variable and the `poor` variable. The following code will help you do this. You can also type `?paste` to learn more about this paste function.

```
age_poor_cat = paste(ageCat, poor, sep=" ")
```



# Medical expenditures by major smoking–caused disease (MSCD) Stratified by age group and poverty status

