NHANES cleaning hints

Stephen Lauer

November 10, 2017

1) Use the `select()` function from the `dplyr` package.

2) Phrased differently, what is the `min`imum age in the data frame that is not `NA` for `BPSys1`? Keep only the people older than this with `filter()` from the `dplyr` package.

3) Here you can use `filter()` twice in opposite ways. Once by using `is.na()` and another time using `!is.na()`. Also, you can use `select()` with the minus sign to remove columns.

4) Inside the `gather()` function, you can use `starts_with()` to select columns that start with a given string.

5) Use the `stringr` package with `str_remove()` on the first part of the `SysMeasure` column. Then use `as.numeric()` to convert the column to numbers.

6) Use `filter(x == y)` to keep values in column `x` that are the same as those in column `y`.

7) More `filter()`ing!

8) Use `group_by()` on the variables you want to keep then `summarise()` with `mean()`.

9) Use `lm()` to fit your linear model and use `plot()` and `summary()` to check the fit.

10) `predict(fit, newdata)`. Then you can use `summary()` to see the distribution of the predictions.

Bonus) `BP_compare <- left_join(final_data_frame, earlier_data_frame)`. Then can do `summary(BP_compare$BPSysA`
to see the distribution of the differences.