

McGill University Department of Epidemiology, Biostatistics and Occupational health

BIOS 640

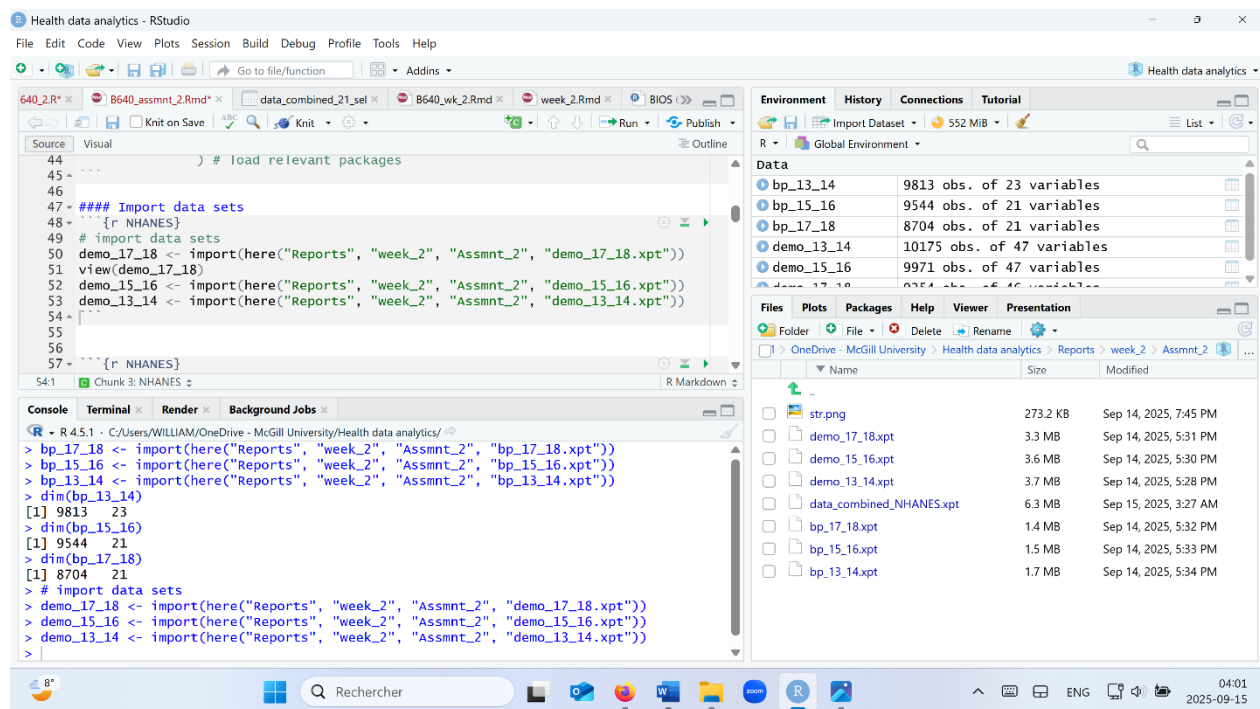
Introduction to Health Data Science Methods

Assessment Week 2

William Opoku-Nkoom

14 September 2025

## Exercise 2: Data import and preliminary exploration



**Fig. 1: View of codes, codes, Console, Environment and directory of imported data sets**

**Table 1: Dimensions of the data sets**

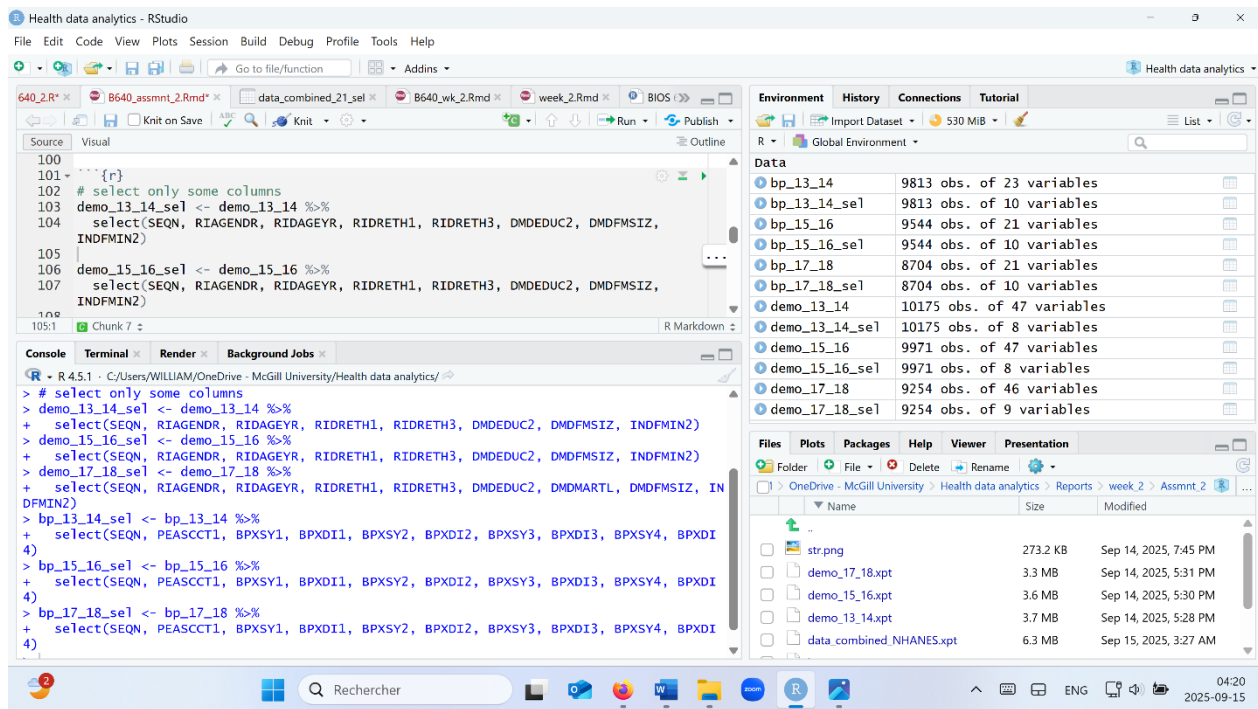
Filename	File description	Dimensions	
		Number of rows	Number of columns
demo_17_18	2017-2018 demographic data	9254	46
demo_15_16	2015-2016 demographic data	9971	47
demo_13_14	2013-2014 demographic data	10175	47
bp_17_18	2017-2018 blood pressure examination data	8704	21
bp_15_16	2015-2016 blood pressure examination data	9544	21
bp_13_14	2013-2014 blood pressure examination data	9813	23

### Types of variables present in the data sets

1. Continuous variable
2. Discrete variables
3. Categorical

All variables are in the right format.

## Exercise 3: Merging and preparing datasets



**Fig. 2: Output of selected columns of the data sets in Console and the Environment Panes of the data sets**

**Table 2: Dimensions of the combined demographic and blood pressure data sets**

Filename	File description	Dimensions	
		Number of rows	Number of columns
data_combined	Combined demographic and blood pressure data for 2013-2104, 2015-2016 and 2017-2018 of the NHANES data sets	29400	20
bp_combined	Combined blood pressure data for 2013-2104, 2015-2016 and 2017-2018 of the NHANES data sets	28061	11
demo_combined	Combined demographic data for 2013-2104, 2015-2016 and 2017-2018 of the NHANES data sets	29400	10

The combined data set from both the combined demographic and blood pressure data have higher dimensions, i.e. it retains all the rows of the combined demographic data set which has higher number of rows compared to the combined blood pressure data set. By this no

respondent data was lost in the combining of these two data sets. Furthermore the combined data set of both has higher number of columns, meaning it retained all columns of the two data sets but overwrote one column which had the same column name, most likely the respondent sequence number.

### Exercise 3: Merging and preparing datasets

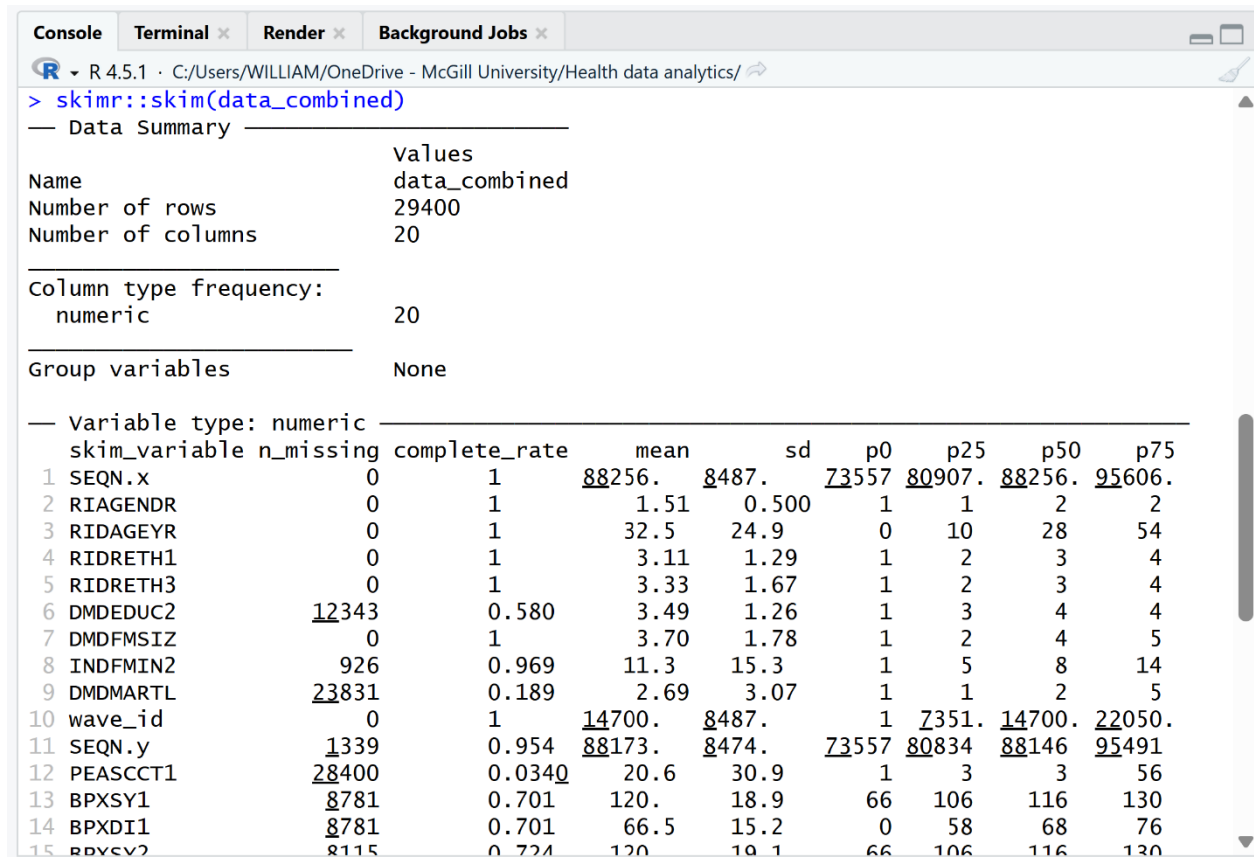
The screenshot shows the RStudio interface with the following components:

- Source Pane:** Contains R code for merging datasets. The code includes comments and functions like `bind_rows`, `full_join`, and `bind` to create `demo_combined` and `bp_combined`, and then merge them into `data_combined` using `full_join` with `by = "wave_id"`.
- Console:** Shows an error message: `Erreur : objet 'demo_combined' introuvable` (Error: object 'demo\_combined' not found). Below the error, the same R code is pasted again.
- Environment Pane:** Lists the objects in the Global Environment. The objects and their dimensions are:
 

Object	Dimensions
bp_15_16	9544 obs. of 21 variables
bp_15_16_sel	9544 obs. of 10 variables
bp_17_18	8704 obs. of 21 variables
bp_17_18_sel	8704 obs. of 10 variables
bp_combined	28061 obs. of 11 variables
data_combined	29400 obs. of 20 variables
demo_13_14	10175 obs. of 47 variables
demo_13_14_sel	10175 obs. of 8 variables
demo_15_16	9971 obs. of 47 variables
demo_15_16_sel	9971 obs. of 8 variables
demo_17_18	9254 obs. of 46 variables
demo_17_18_sel	9254 obs. of 9 variables
demo_combined	29400 obs. of 10 variables
- Files Pane:** Shows a list of files in the current directory, including `str.png`, `demo_17_18.xpt`, `demo_15_16.xpt`, `demo_13_14.xpt`, and `data_combined_NHANES.xpt`.

**Fig. 3: Outputs of combined data sets in the Console and the Environment Panes**

## Exercise 4: Data cleaning and variable recoding



```
R 4.5.1 · C:/Users/WILLIAM/OneDrive - McGill University/Health data analytics/
> skimr::skim(data_combined)
```

— Data Summary —

Name	data_combined
Number of rows	29400
Number of columns	20
Column type frequency:	
numeric	20
Group variables	None

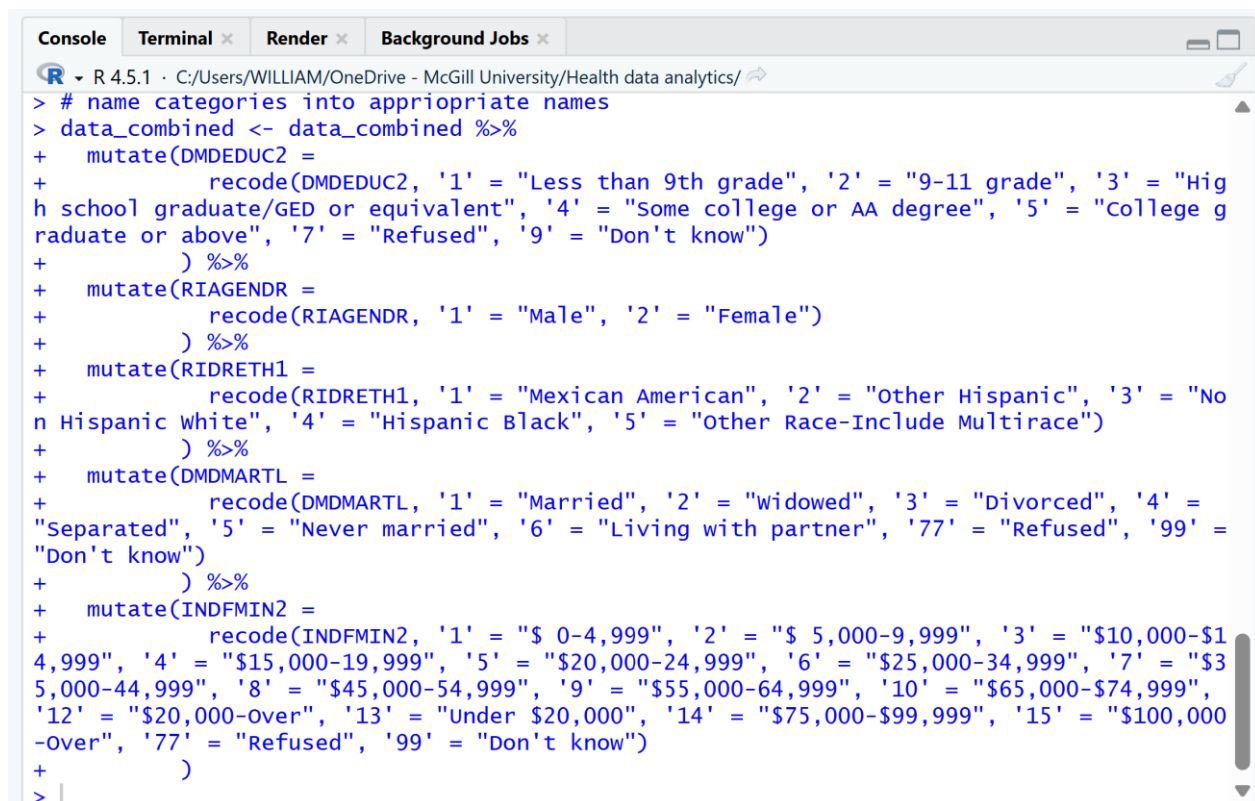
— Variable type: numeric —

	skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75
1	SEQN.x	0	1	88256.	8487.	73557	80907.	88256.	95606.
2	RIAGENDR	0	1	1.51	0.500	1	1	2	2
3	RIDAGEYR	0	1	32.5	24.9	0	10	28	54
4	RIDRETH1	0	1	3.11	1.29	1	2	3	4
5	RIDRETH3	0	1	3.33	1.67	1	2	3	4
6	DMDEDUC2	12343	0.580	3.49	1.26	1	3	4	4
7	DMDFMSIZ	0	1	3.70	1.78	1	2	4	5
8	INDFMIN2	926	0.969	11.3	15.3	1	5	8	14
9	DMDMARTL	23831	0.189	2.69	3.07	1	1	2	5
10	wave_id	0	1	14700.	8487.	1	7351.	14700.	22050.
11	SEQN.y	1339	0.954	88173.	8474.	73557	80834	88146	95491
12	PEASCCT1	28400	0.0340	20.6	30.9	1	3	3	56
13	BPXSY1	8781	0.701	120.	18.9	66	106	116	130
14	BPXD11	8781	0.701	66.5	15.2	0	58	68	76
15	BPXSY2	8115	0.724	120.	19.1	66	106	116	130

**Fig. 4: Skim data output in the Console Pane**

Skim() function observation: there were a lot of missing values for the following variables:

DMDEDUC2, INDFMIN2, DMDMARTL, SEQN.y, PEASCCT1, BPXSY1, BPXD11, BPXSY2, BPXD12, BPXSY3, BPXD13, BPXSY4, BPXD14.

The image shows a screenshot of an R console window. The window has a title bar with tabs for 'Console', 'Terminal', 'Render', and 'Background Jobs'. The 'Console' tab is active. The R version is 4.5.1, and the working directory is C:/Users/WILLIAM/OneDrive - McGill University/Health data analytics/. The code in the console is as follows:

```
> # name categories into appropriate names
> data_combined <- data_combined %>%
+   mutate(DMDEDUC2 =
+     recode(DMDEDUC2, '1' = "Less than 9th grade", '2' = "9-11 grade", '3' = "High school graduate/GED or equivalent", '4' = "Some college or AA degree", '5' = "College graduate or above", '7' = "Refused", '9' = "Don't know")
+   ) %>%
+   mutate(RIAGENDR =
+     recode(RIAGENDR, '1' = "Male", '2' = "Female")
+   ) %>%
+   mutate(RIDRETH1 =
+     recode(RIDRETH1, '1' = "Mexican American", '2' = "Other Hispanic", '3' = "Non Hispanic White", '4' = "Hispanic Black", '5' = "Other Race-Include Multirace")
+   ) %>%
+   mutate(DMDMARTL =
+     recode(DMDMARTL, '1' = "Married", '2' = "Widowed", '3' = "Divorced", '4' = "Separated", '5' = "Never married", '6' = "Living with partner", '77' = "Refused", '99' = "Don't know")
+   ) %>%
+   mutate(INDFMIN2 =
+     recode(INDFMIN2, '1' = "$ 0-4,999", '2' = "$ 5,000-9,999", '3' = "$10,000-$14,999", '4' = "$15,000-19,999", '5' = "$20,000-24,999", '6' = "$25,000-34,999", '7' = "$35,000-44,999", '8' = "$45,000-54,999", '9' = "$55,000-64,999", '10' = "$65,000-$74,999", '12' = "$20,000-Over", '13' = "Under $20,000", '14' = "$75,000-$99,999", '15' = "$100,000-Over", '77' = "Refused", '99' = "Don't know")
+   )
>
```

**Fig. 5: Output of some recoded variables in the Console Pane**

## Exercise 5: Filtering and reshaping a dataset

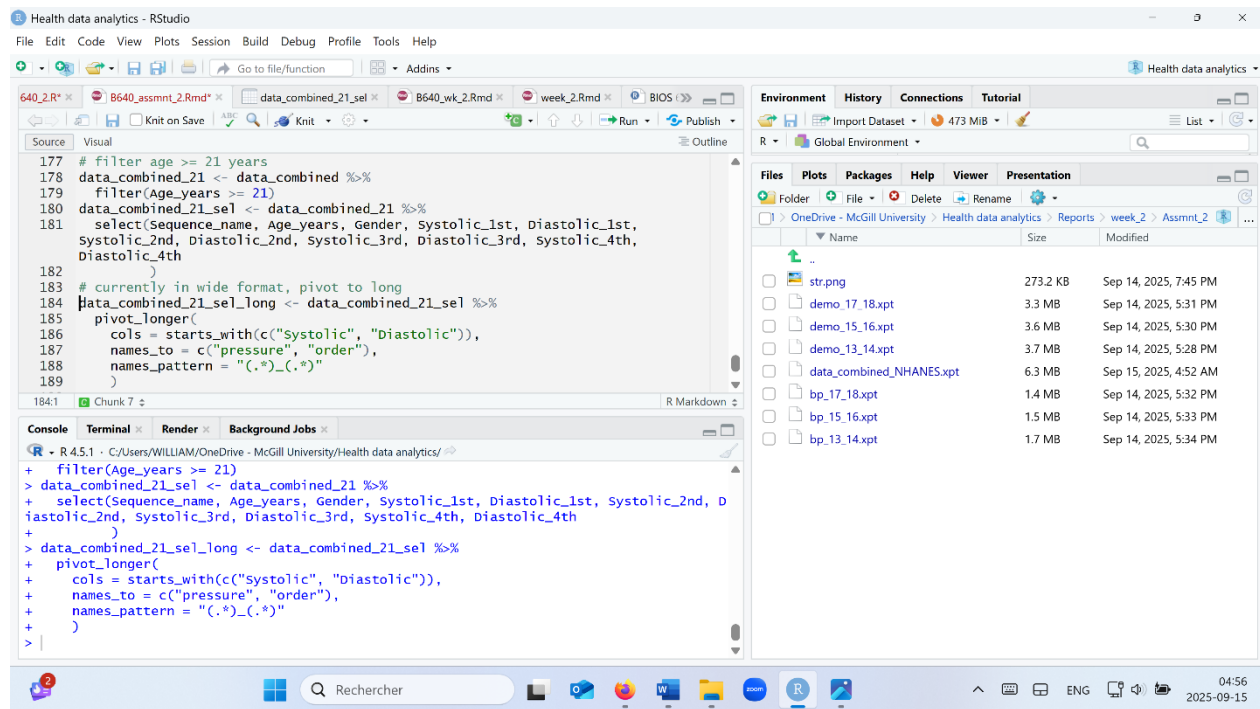


Fig. 6: Output of filtered, selected and pivoted data in Console Pane

The resulting data sets are in the wider format because the systolic and the diastolic blood pressure values 1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, and 4<sup>th</sup> readings are all in different columns. The data was pivoted into the longer format as shown in Fig. 6.