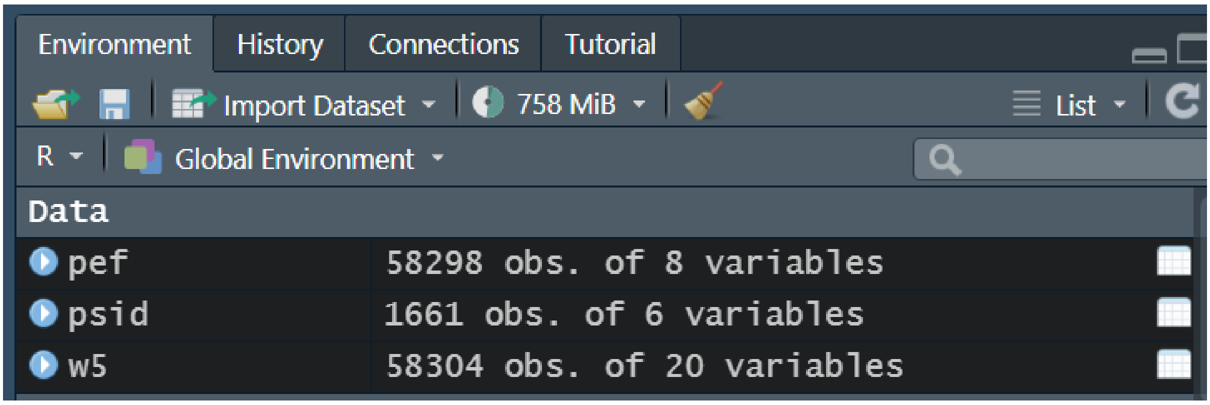
Tugas Biostatistik Lanjut

Mahasiswa: dr. Ezra Hans Soputra
NIM: 131520230501
Dosen Pengajar: Dwi Agustian, dr., MPH, PhD

Soal:
Mempersiapkan data untuk analisis variabel pef, age, sex, height:
1. Membuat kode untuk membersihkan (mengexclude) missing data
2. Membuat data profile diagram yang berisikan langkah - langkah penyiapan data sejak file pef & w5 baca/extraks sampai dengan data yang siap dianalisis. dengan jumlah obervasi /records/subjects dari tiap-tiap langkah tertuliskan pada diagram tersebut.
Kode dan data profile diagram dalam bentuk pdf/png dapat diupload di Git hub dan link dari masing-masingnya diposting di submission tugas ini.

1. #mengimport data dari URL

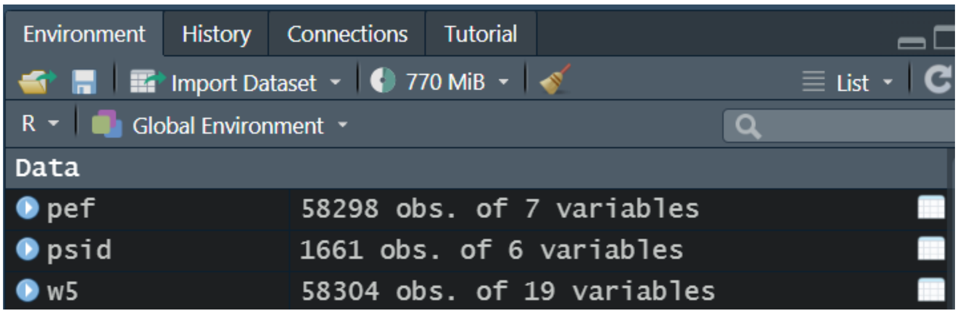
```
library(readr)
pef <- read_csv("https://raw.githubusercontent.com/dwi-agustian/biostat/main/pef.csv")

w5 <- read_csv("https://raw.githubusercontent.com/dwi-agustian/biostat/main/w5.csv")

library(dplyr)
```



```
Data
pef     58298 obs. of 8 variables
psid    1661 obs. of 6 variables
w5      58304 obs. of 20 variables
```

2. #memilih variable yang akan dianalisis

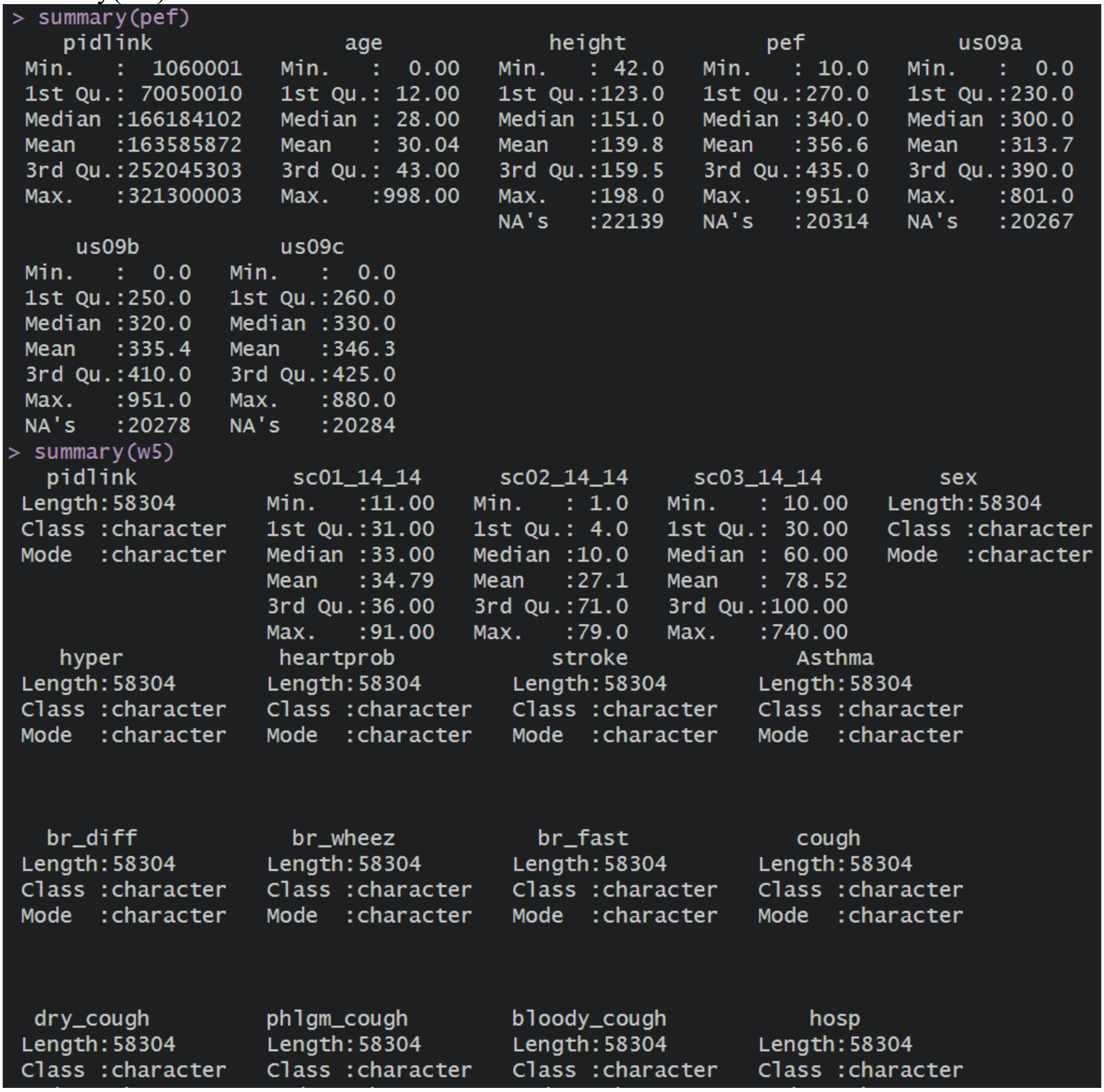
```
pef = select(pef, pidlink,age,height,pef,us09a,us09b,us09c)

w5 = select(w5,
pidlink,sc01_14_14,sc02_14_14,sc03_14_14,sex,hyper,heartprob,stroke,Asthma,br_diff,br_wheez,br_fast,cough,dry_cough,phlgm_cough,bloody_cough,hosp,outp,agegr)
```



```
Data
pef     58298 obs. of 7 variables
psid    1661 obs. of 6 variables
w5      58304 obs. of 19 variables
```

3. #melihat deskriptif statistik dari semua variable di data set

summary(pef)
summary(w5)

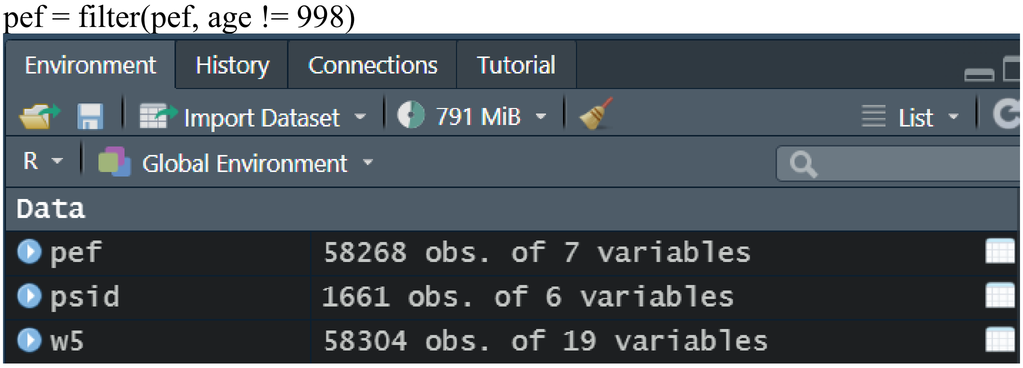


```
> summary(pef)
    pidlink              age             height          pef             us09a
 Min.   :  1060001   Min.   :  0.00   Min.   : 42.0   Min.   : 10.0   Min.   :  0.0
 1st Qu.: 70050010   1st Qu.: 12.00   1st Qu.:123.0   1st Qu.:270.0   1st Qu.:230.0
 Median :166184102   Median : 28.00   Median :151.0   Median :340.0   Median :300.0
 Mean   :163585872   Mean   : 30.04   Mean   :139.8   Mean   :356.6   Mean   :313.7
 3rd Qu.:252045303   3rd Qu.: 43.00   3rd Qu.:159.5   3rd Qu.:435.0   3rd Qu.:390.0
 Max.   :321300003   Max.   :998.00   Max.   :198.0   Max.   :951.0   Max.   :801.0
                                      NA's   :22139   NA's   :20314   NA's   :20267
     us09b            us09c
 Min.   :  0.0   Min.   :  0.0
 1st Qu.:250.0   1st Qu.:260.0
 Median :320.0   Median :330.0
 Mean   :335.4   Mean   :346.3
 3rd Qu.:410.0   3rd Qu.:425.0
 Max.   :951.0   Max.   :880.0
 NA's   :20278   NA's   :20284
> summary(w5)
   pidlink            sc01_14_14      sc02_14_14     sc03_14_14          sex
 Length:58304       Min.   :11.00   Min.   : 1.0   Min.   : 10.00   Length:58304
 Class :character   1st Qu.:31.00   1st Qu.: 4.0   1st Qu.: 30.00   Class :character
 Mode  :character   Median :33.00   Median :10.0   Median : 60.00   Mode  :character
                    Mean   :34.79   Mean   :27.1   Mean   : 78.52
                    3rd Qu.:36.00   3rd Qu.:71.0   3rd Qu.:100.00
                    Max.   :91.00   Max.   :79.0   Max.   :740.00
    hyper            heartprob            stroke              Asthma
 Length:58304       Length:58304       Length:58304       Length:58304
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



   br_diff            br_wheez            br_fast              cough
 Length:58304       Length:58304       Length:58304       Length:58304
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



  dry_cough         phlgm_cough        bloody_cough           hosp
 Length:58304       Length:58304       Length:58304       Length:58304
 Class :character   Class :character   Class :character   Class :character
```

4. memilih observasi

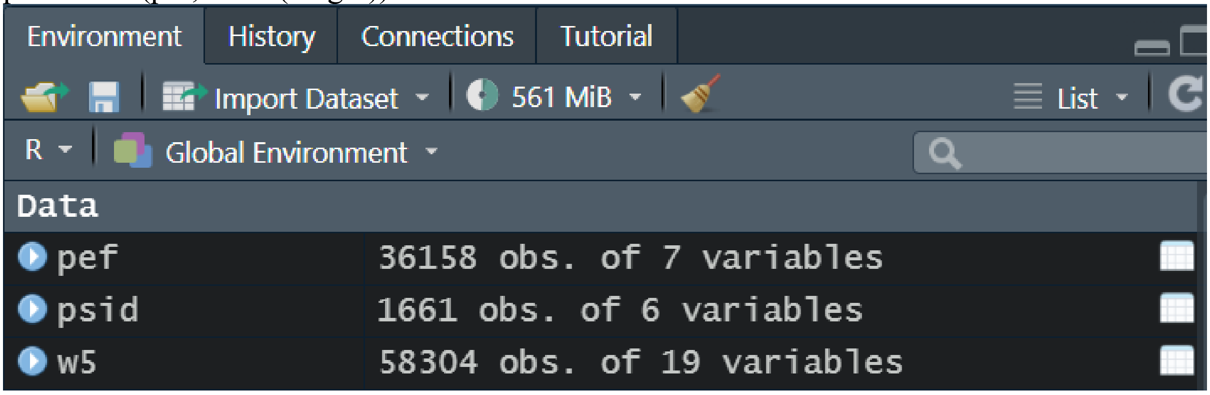
#memilih observasi dimana data height tidak missing (komplit), dan data Age tidak ada yang outlier

pef = filter(pef, age != 998)



```
Data
pef      58268 obs. of 7 variables
psid     1661 obs. of 6 variables
w5       58304 obs. of 19 variables
```

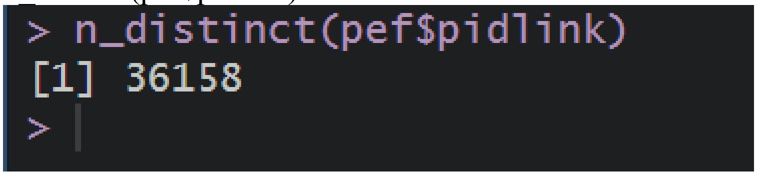
pef = filter(pef, !is.na(height))



```
Data
pef     36158 obs. of 7 variables
psid    1661 obs. of 6 variables
w5      58304 obs. of 19 variables
```

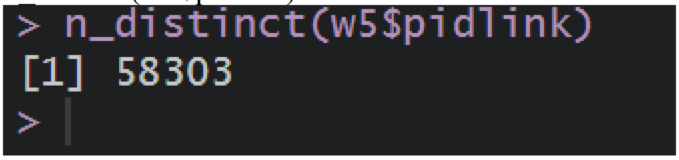
5. # check jumlah obs pidlink yang unik

n_distinct(pef$pidlink)



```
> n_distinct(pef$pidlink)
[1] 36158
> 
```

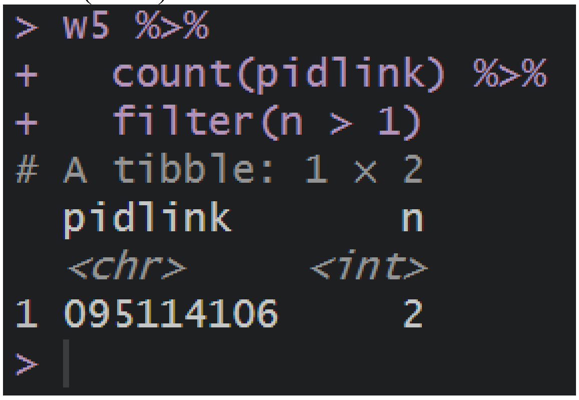
n_distinct(w5$pidlink)



```
> n_distinct(w5$pidlink)
[1] 58303
> 
```

6. #Find duplicated Pidlink

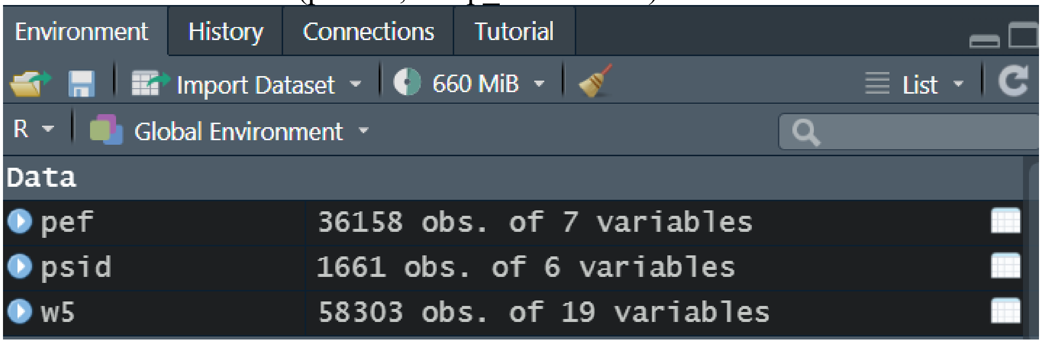
```
w5 %>%
  count(pidlink) %>%
  filter(n > 1)
```



```
> w5 %>%
+   count(pidlink) %>%
+   filter(n > 1)
# A tibble: 1 × 2
  pidlink       n
  <chr>     <int>
1 095114106     2
> 
```

7. # Remove duplicated rows based on pidlink

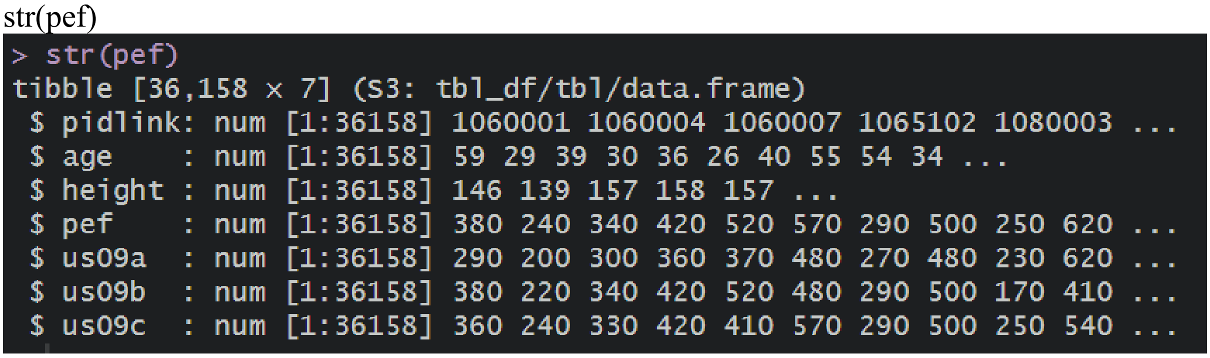
w5 = w5 %>% distinct(pidlink, .keep_all = TRUE)



```
Data
pef                 36158 obs. of 7 variables
psid                1661 obs. of 6 variables
w5                  58303 obs. of 19 variables
```

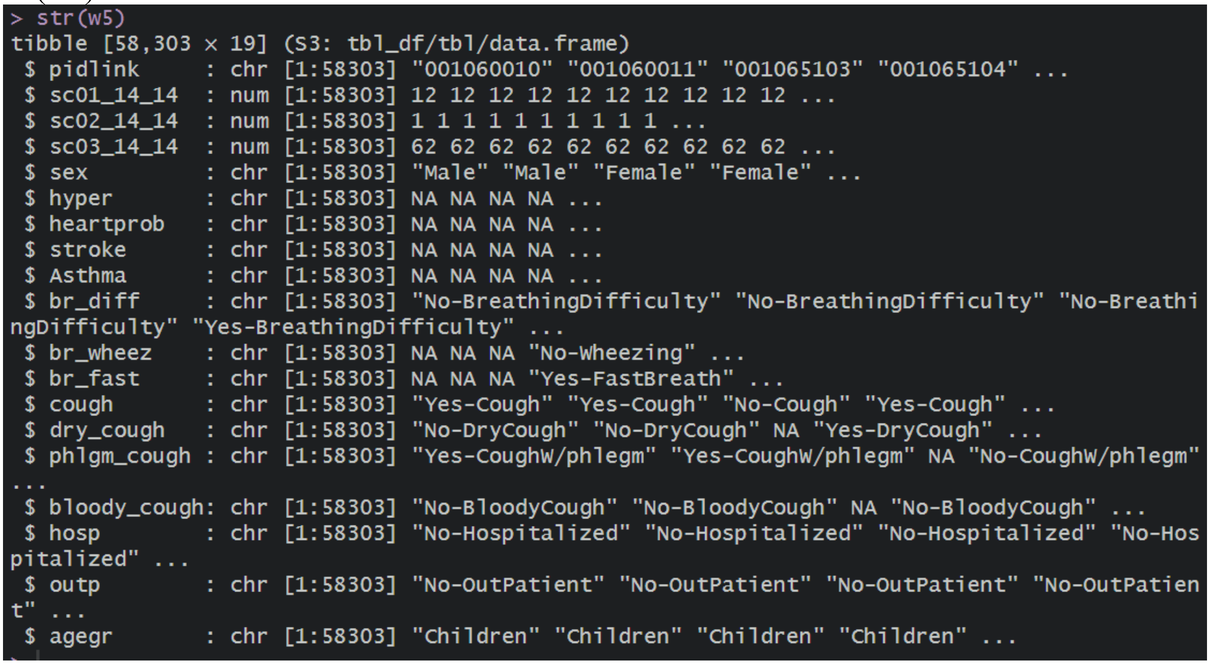
8. #melakukan check klasifikasi variabel

str(pef)



```
> str(pef)
tibble [36,158 × 7] (S3: tbl_df/tbl/data.frame)
 $ pidlink: num [1:36158] 1060001 1060004 1060007 1065102 1080003 ...
 $ age    : num [1:36158] 59 29 39 30 36 26 40 55 54 34 ...
 $ height : num [1:36158] 146 139 157 158 157 ...
 $ pef    : num [1:36158] 380 240 340 420 520 570 290 500 250 620 ...
 $ us09a  : num [1:36158] 290 200 300 360 370 480 270 480 230 620 ...
 $ us09b  : num [1:36158] 380 220 340 420 520 480 290 500 170 410 ...
 $ us09c  : num [1:36158] 360 240 330 420 410 570 290 500 250 540 ...
```

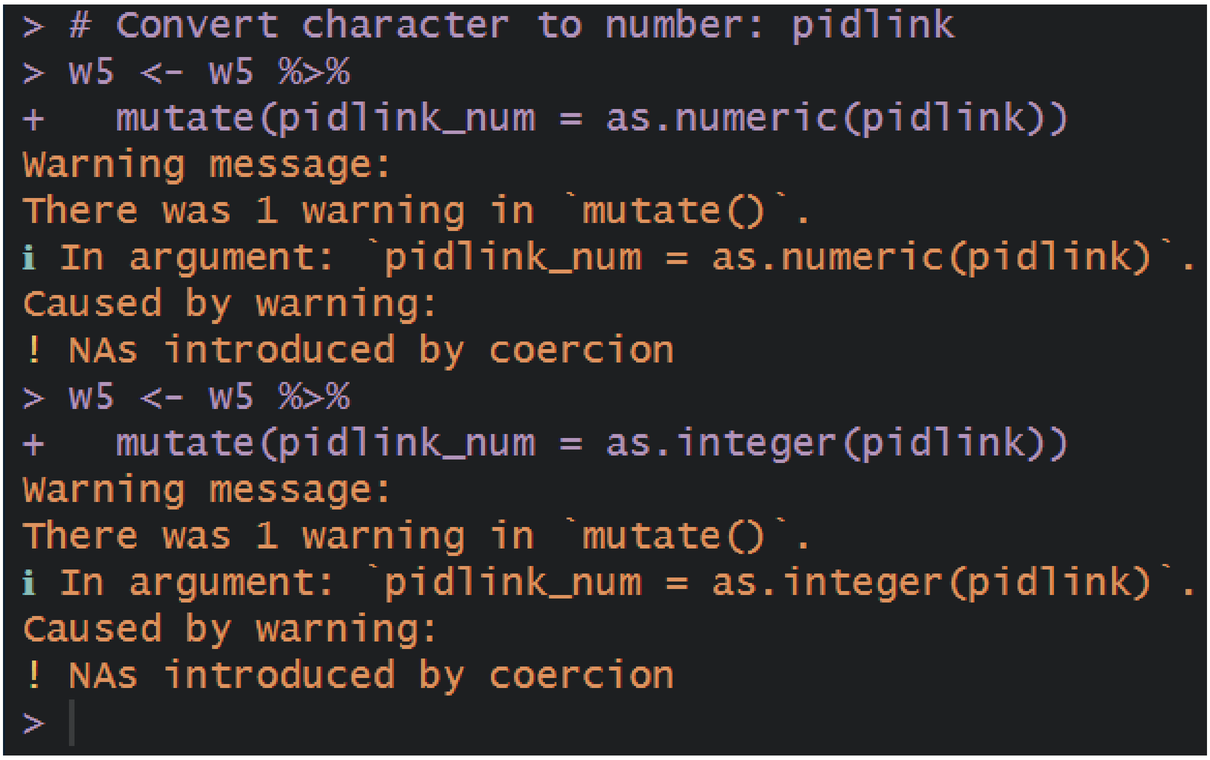
str(w5)



```
> str(w5)
tibble [58,303 × 19] (S3: tbl_df/tbl/data.frame)
 $ pidlink      : chr [1:58303] "001060010" "001060011" "001065103" "001065104" ...
 $ sc01_14_14   : num [1:58303] 12 12 12 12 12 12 12 12 12 12 ...
 $ sc02_14_14   : num [1:58303] 1 1 1 1 1 1 1 1 1 1 ...
 $ sc03_14_14   : num [1:58303] 62 62 62 62 62 62 62 62 62 62 ...
 $ sex          : chr [1:58303] "Male" "Male" "Female" "Female" ...
 $ hyper        : chr [1:58303] NA NA NA NA ...
 $ heartprob    : chr [1:58303] NA NA NA NA ...
 $ stroke       : chr [1:58303] NA NA NA NA ...
 $ Asthma       : chr [1:58303] NA NA NA NA ...
 $ br_diff      : chr [1:58303] "No-BreathingDifficulty" "No-BreathingDifficulty" "No-Breathi
ngDifficulty" "Yes-BreathingDifficulty" ...
 $ br_wheez     : chr [1:58303] NA NA NA "No-Wheezing" ...
 $ br_fast      : chr [1:58303] NA NA NA "Yes-FastBreath" ...
 $ cough        : chr [1:58303] "Yes-Cough" "Yes-Cough" "No-Cough" "Yes-Cough" ...
 $ dry_cough    : chr [1:58303] "No-DryCough" "No-DryCough" NA "Yes-DryCough" ...
 $ phlgm_cough  : chr [1:58303] "Yes-CoughW/phlegm" "Yes-CoughW/phlegm" NA "No-CoughW/phlegm"
...
 $ bloody_cough: chr [1:58303] "No-BloodyCough" "No-BloodyCough" NA "No-BloodyCough" ...
 $ hosp         : chr [1:58303] "No-Hospitalized" "No-Hospitalized" "No-Hospitalized" "No-Hos
pitalized" ...
 $ outp         : chr [1:58303] "No-OutPatient" "No-OutPatient" "No-OutPatient" "No-OutPatien
t" ...
 $ agegr        : chr [1:58303] "Children" "Children" "Children" "Children" ...
```

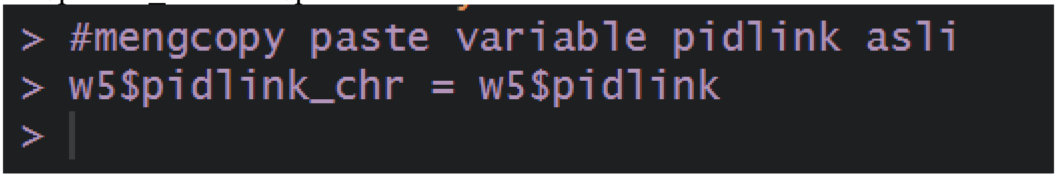
9. # Convert character to number: pidlink

```
w5 <- w5 %>%
  mutate(pidlink_num = as.numeric(pidlink))

w5 <- w5 %>%
  mutate(pidlink_num = as.integer(pidlink))
```



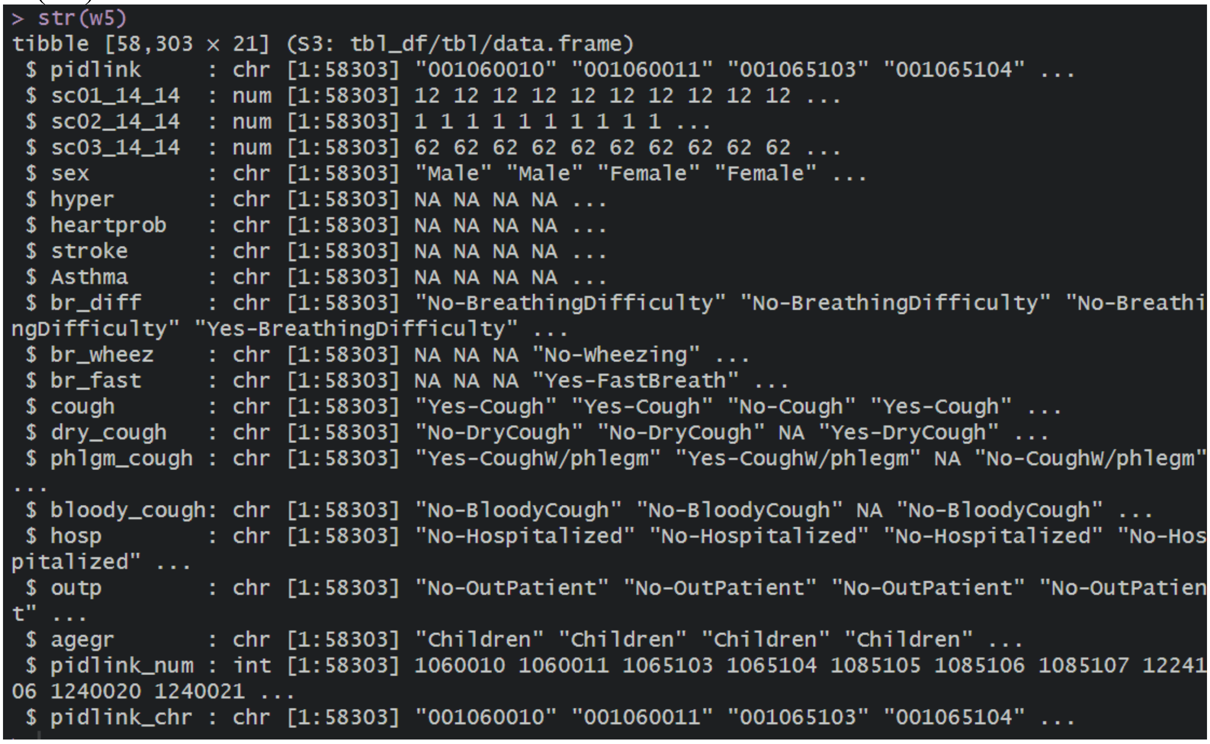
```
> # Convert character to number: pidlink
> w5 <- w5 %>%
+   mutate(pidlink_num = as.numeric(pidlink))
Warning message:
There was 1 warning in `mutate()`.
ℹ In argument: `pidlink_num = as.numeric(pidlink)`.
Caused by warning:
! NAs introduced by coercion
> w5 <- w5 %>%
+   mutate(pidlink_num = as.integer(pidlink))
Warning message:
There was 1 warning in `mutate()`.
ℹ In argument: `pidlink_num = as.integer(pidlink)`.
Caused by warning:
! NAs introduced by coercion
> 
```

10. #mengcopy paste variable pidlink asli
w5$pidlink_chr = w5$pidlink



```
> #mengcopy paste variable pidlink asli
> w5$pidlink_chr = w5$pidlink
> 
```

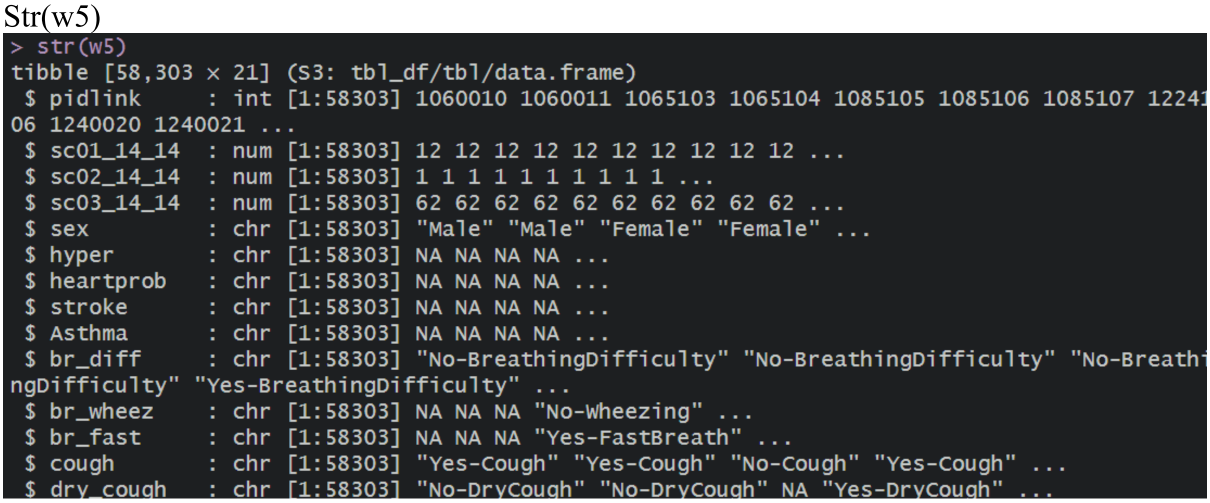
11. #melakukan check klasifikasi variabel
str(w5)



```
> str(w5)
tibble [58,303 × 21] (S3: tbl_df/tbl/data.frame)
 $ pidlink     : chr [1:58303] "001060010" "001060011" "001065103" "001065104" ...
 $ sc01_14_14  : num [1:58303] 12 12 12 12 12 12 12 12 12 12 ...
 $ sc02_14_14  : num [1:58303] 1 1 1 1 1 1 1 1 1 1 ...
 $ sc03_14_14  : num [1:58303] 62 62 62 62 62 62 62 62 62 62 ...
 $ sex         : chr [1:58303] "Male" "Male" "Female" "Female" ...
 $ hyper       : chr [1:58303] NA NA NA NA ...
 $ heartprob   : chr [1:58303] NA NA NA NA ...
 $ stroke      : chr [1:58303] NA NA NA NA ...
 $ Asthma      : chr [1:58303] NA NA NA NA ...
 $ br_diff     : chr [1:58303] "No-BreathingDifficulty" "No-BreathingDifficulty" "No-Breathi
ngDifficulty" "Yes-BreathingDifficulty" ...
 $ br_wheez    : chr [1:58303] NA NA NA "No-Wheezing" ...
 $ br_fast     : chr [1:58303] NA NA NA "Yes-FastBreath" ...
 $ cough       : chr [1:58303] "Yes-Cough" "Yes-Cough" "No-Cough" "Yes-Cough" ...
 $ dry_cough   : chr [1:58303] "No-DryCough" "No-DryCough" NA "Yes-DryCough" ...
 $ phlgm_cough : chr [1:58303] "Yes-CoughW/phlegm" "Yes-CoughW/phlegm" NA "No-CoughW/phlegm"
...
 $ bloody_cough: chr [1:58303] "No-BloodyCough" "No-BloodyCough" NA "No-BloodyCough" ...
 $ hosp        : chr [1:58303] "No-Hospitalized" "No-Hospitalized" "No-Hospitalized" "No-Hos
pitalized" ...
 $ outp        : chr [1:58303] "No-OutPatient" "No-OutPatient" "No-OutPatient" "No-OutPatien
t" ...
 $ agegr       : chr [1:58303] "Children" "Children" "Children" "Children" ...
 $ pidlink_num : int [1:58303] 1060010 1060011 1065103 1065104 1085105 1085106 1085107 12241
06 1240020 1240021 ...
 $ pidlink_chr : chr [1:58303] "001060010" "001060011" "001065103" "001065104" ...
```

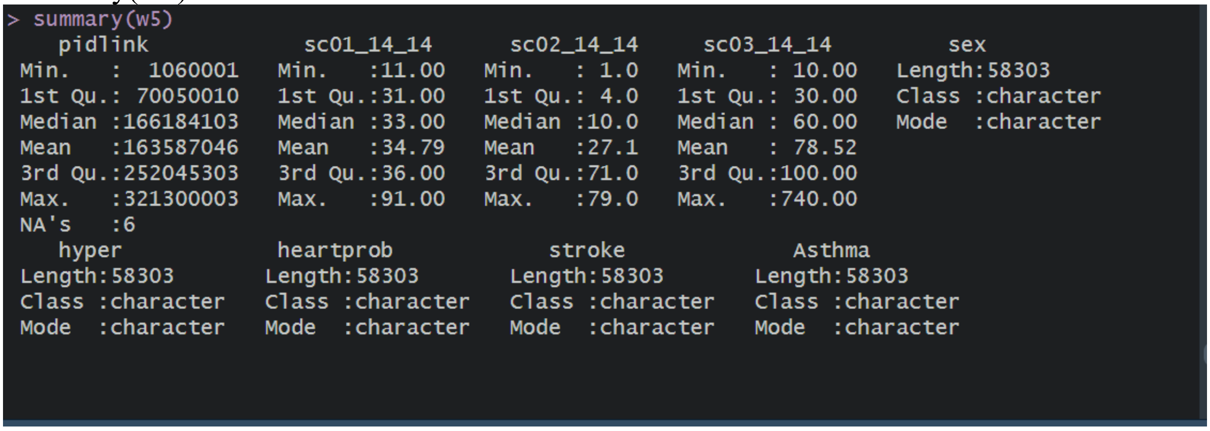
12. #mereplace pidlink dengan versi int(num)
w5$pidlink = w5$pidlink_num

13. #melakukan check klasifikasi variabel w5
Str(w5)



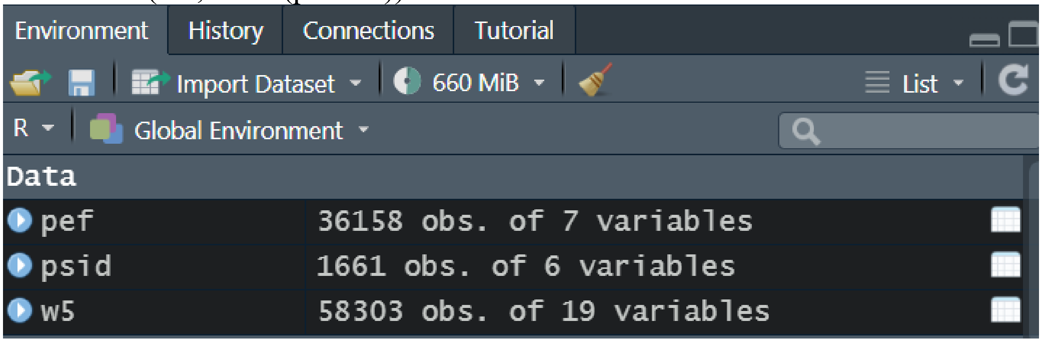
```
> str(w5)
tibble [58,303 × 21] (S3: tbl_df/tbl/data.frame)
 $ pidlink     : int [1:58303] 1060010 1060011 1065103 1065104 1085105 1085106 1085107 12241
06 1240020 1240021 ...
 $ sc01_14_14  : num [1:58303] 12 12 12 12 12 12 12 12 12 12 ...
 $ sc02_14_14  : num [1:58303] 1 1 1 1 1 1 1 1 1 1 ...
 $ sc03_14_14  : num [1:58303] 62 62 62 62 62 62 62 62 62 62 ...
 $ sex         : chr [1:58303] "Male" "Male" "Female" "Female" ...
 $ hyper       : chr [1:58303] NA NA NA NA ...
 $ heartprob   : chr [1:58303] NA NA NA NA ...
 $ stroke      : chr [1:58303] NA NA NA NA ...
 $ Asthma      : chr [1:58303] NA NA NA NA ...
 $ br_diff     : chr [1:58303] "No-BreathingDifficulty" "No-BreathingDifficulty" "No-Breathi
ngDifficulty" "Yes-BreathingDifficulty" ...
 $ br_wheez    : chr [1:58303] NA NA NA "No-Wheezing" ...
 $ br_fast     : chr [1:58303] NA NA NA "Yes-FastBreath" ...
 $ cough       : chr [1:58303] "Yes-Cough" "Yes-Cough" "No-Cough" "Yes-Cough" ...
 $ dry_cough   : chr [1:58303] "No-DryCough" "No-DryCough" NA "Yes-DryCough" ...
```

14. #melihat deskriptif statistik dari semua variable di data set w5
Summary(w5)

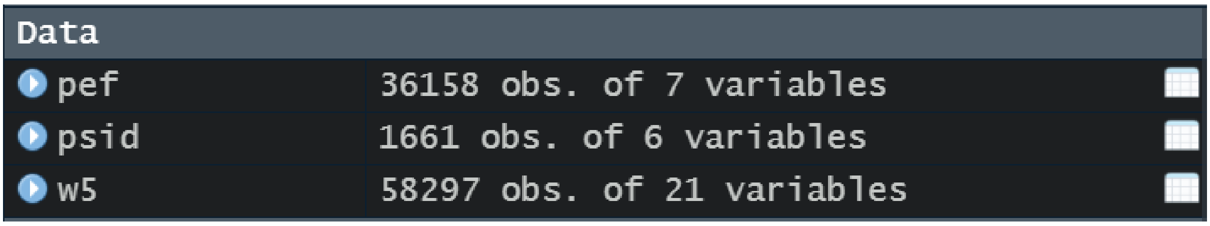


```
> summary(w5)
    pidlink             sc01_14_14       sc02_14_14      sc03_14_14          sex           
 Min.   :  1060001   Min.   :11.00   Min.   : 1.0   Min.   : 10.00   Length:58303      
 1st Qu.: 70050010   1st Qu.:31.00   1st Qu.: 4.0   1st Qu.: 30.00   Class :character  
 Median :166184103   Median :33.00   Median :10.0   Median : 60.00   Mode  :character  
 Mean   :163587046   Mean   :34.79   Mean   :27.1   Mean   : 78.52                     
 3rd Qu.:252045303   3rd Qu.:36.00   3rd Qu.:71.0   3rd Qu.:100.00                     
 Max.   :321300003   Max.   :91.00   Max.   :79.0   Max.   :740.00                     
 NA's   :6                                                                             
    hyper             heartprob            stroke              Asthma         
 Length:58303       Length:58303       Length:58303       Length:58303      
 Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character  
```

15. #memilih observasi dimana pidlink w5 tidak missing
w5 = filter (w5, !is.na(pidlink))



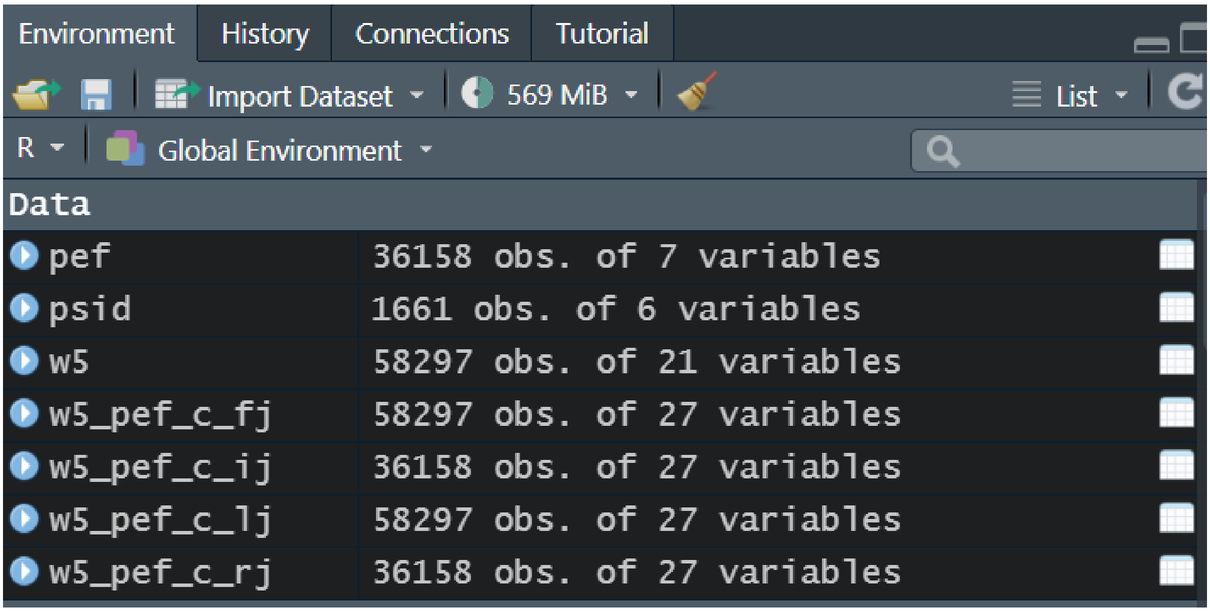
```
Data
pef     36158 obs. of 7 variables
psid    1661 obs. of 6 variables
w5      58303 obs. of 19 variables
```



```
Data
pef     36158 obs. of 7 variables
psid    1661 obs. of 6 variables
w5      58297 obs. of 21 variables
```

16. #combining dataset (menggabungkan data)

```
# menggabungkan variables dengan common variable
w5_pef_c_lj = left_join(w5, pef, by = "pidlink")
w5_pef_c_rj = right_join(w5, pef, by = "pidlink")
w5_pef_c_ij = inner_join(w5, pef, by = "pidlink")
w5_pef_c_fj = full_join(w5, pef, by = "pidlink")
```



```
Data
pef            36158 obs. of 7 variables
psid           1661 obs. of 6 variables
w5             58297 obs. of 21 variables
w5_pef_c_fj    58297 obs. of 27 variables
w5_pef_c_ij    36158 obs. of 27 variables
w5_pef_c_lj    58297 obs. of 27 variables
w5_pef_c_rj    36158 obs. of 27 variables
```