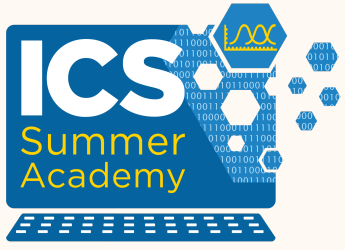

Data Wrangling in R

Zahra Moslemi

Adapted from slides by Mine Dogucu

```
glimpse(AD)
```

```
## Rows: 2,700
## Columns: 57
## $ id        <chr> "S060833", "S932623", "S755478", "S852291"
## $ diagnosis <fct> Normal cognition, Normal cognition, Normal
## $ age       <dbl> 74, 56, 77, 74, 75, 72, 64, 78, 73, 81, 66
## $ educ      <dbl> 12, 16, 18, 20, 14, 16, 16, 17, 18, 13, 16
## $ female    <fct> male, female, female, female, male, female
## $ height    <dbl> 65.0, 62.0, 65.0, 62.0, 62.0, 61.8, 60.0,
## $ weight    <dbl> 233, 110, 137, 112, 127, 141, 124, 152, 13
## $ bpsys     <dbl> 148, 110, 144, 120, 145, 107, 112, 134, 12
## $ bpdias    <dbl> 100, 75, 60, 60, 61, 65, 70, 74, 60, 70, 8
## $ hrate     <dbl> 72, 60, 64, 72, 58, 83, 76, 70, 60, 76, 60
## $ cdrglob   <dbl> 0.5, 0.0, 0.0, 0.0, 0.5, 0.0, 0.0, 0.5, 0.
## $ naccgds   <dbl> 5, 1, 0, 0, 4, 1, 2, 0, 0, 5, 0, 1, 0, 0,
## $ delsev    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
```

```
colnames(AD)
```

```
##  [1] "id"        "diagnosis" "age"       "educ"      "female
##  [7] "weight"    "bpsys"     "bpdias"    "hrate"     "cdrglo
## [13] "delsev"    "hallsev"   "agitsev"   "depdsev"   "anxsev
## [19] "apasev"    "disnsev"   "irrsev"    "motsev"    "nitese
## [25] "bills"     "taxes"     "shopping"  "games"     "stove"
## [31] "events"    "payattn"   "remdates"  "travel"    "naccmm
## [37] "digif"     "animals"   "traila"    "trailb"    "naccic
## [43] "lhippo"    "rhippo"    "frcort"    "lparcort"  "rparco
## [49] "rtempcor"  "lcac"      "rcac"      "lent"      "rent"
## [55] "rparhip"   "lposcin"   "rposcin"
```

subsetting variables/columns

	variable_1	variable_2	variable_3	variable_4
1				
2				
3				
4				

	variable_2	variable_3
1		
2		
3		
4		

`select()`

subsetting observations/rows

	variable_1	variable_2	variable_3	variable_4
1				
2				
3				
4				

	variable_1	variable_2	variable_3	variable_4
1				
2				

`slice()` and `filter()`

`select` is used to select certain variables in the data frame.

```
select(AD, age, cdrglob)
```

```
## # A tibble: 2,700 × 2
##      age cdrglob
##    <dbl>   <dbl>
##  1    74     0.5
##  2    56     0
##  3    77     0
##  4    74     0
##  5    75     0.5
##  6    72     0
##  7    64     0
##  8    78     0.5
##  9    73     0
## 10    81     1
```

```
AD %>%
  select(age, cdrglob)
```

```
## # A tibble: 2,700 × 2
##      age cdrglob
##    <dbl>   <dbl>
##  1    74     0.5
##  2    56     0
##  3    77     0
##  4    74     0
##  5    75     0.5
##  6    72     0
##  7    64     0
##  8    78     0.5
##  9    73     0
## 10    81     1
```

`select` can also be used to drop certain variables if used with a negative sign.

```
select(AD, -id, -female)
```

```
## # A tibble: 2,700 × 55
##    diagnosis   age  educ height weight bpsys bpdias hrate cd
##    <fct>     <dbl> <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl>
##  1 Normal c…    74    12   65      233   148    100    72
##  2 Normal c…    56    16   62      110   110     75    60
##  3 Normal c…    77    18   65      137   144     60    64
##  4 Normal c…    74    20   62      112   120     60    72
##  5 Mild cog…    75    14   62      127   145     61    58
##  6 Normal c…    72    16   61.8    141   107     65    83
##  7 Normal c…    64    16   60      124   112     70    76
##  8 Dementia…    78    17   69      152   134     74    70
##  9 Normal c…    73    18   65      131   122     60    60
## 10 Dementia…    81    13   71      197   120     70    76
```

Selection helpers

`starts_with()`
`ends_with()`
`contains()`

```
select(AD, starts_with("cdrglob"))
```

```
## # A tibble: 2,700 × 1
##    cdrglob
##      <dbl>
##  1     0.5
##  2     0
##  3     0
##  4     0
##  5     0.5
```

```
select(AD, contains("ght"))
```

```
## # A tibble: 2,700 × 2
##    height weight
##     <dbl>  <dbl>
##  1   65      233
##  2   62      110
##  3   65      137
##  4   62      112
##  5   62      127
##  6   61.8    141
##  7   60      124
##  8   69      152
##  9   65      131
## 10   71      197
## # ℹ 2,690 more rows
```

subsetting variables/columns

	variable_1	variable_2	variable_3	variable_4
1				
2				
3				
4				

	variable_2	variable_3
1		
2		
3		
4		

`select()`

subsetting observations/rows

	variable_1	variable_2	variable_3	variable_4
1				
2				
3				
4				

	variable_1	variable_2	variable_3	variable_4
1				
2				

`slice()` and `filter()`

`slice()` subsetting rows based on a row number.

The data below include all the rows from third to seventh. Including third and seventh.

```
slice(AD, 3:7)
```

```
## # A tibble: 5 × 57
##   id      diagnosis    age
##   <chr>   <fct>      <dbl>
## 1 S755478 Normal co…    77
## 2 S852291 Normal co…    74
## 3 S011143 Mild cogn…    75
## 4 S069106 Normal co…    72
## 5 S283729 Normal co…    64
## # ℹ 46 more variables: pasc
```

`filter()` subsetting rows based on a condition.

The data below includes rows when the age is 90.

```
filter(AD, age == 90)
```

```
## # A tibble: 8 × 57
##   id      diagnosis    age
##   <chr>   <fct>      <dbl>
## 1 S600123 Dementia …    90
## 2 S203848 Dementia …    90
## 3 S687424 Normal co…    90
## 4 S953670 Dementia …    90
## 5 S146311 Normal co…    90
## 6 S514308 Normal co…    90
## 7 S979193 Dementia …    90
```

Relational Operators in R

Operator	Description
<	Less than
>	Greater than
<=	Less than or equal to
>=	Greater than or equal to
==	Equal to
!=	Not equal to

Logical Operators in R

Operator	Description
&	and
|	or

Recall that when cdrglob (Global Clinical Dementia Rating (CDR) Score) == 3.0 it was identified as severe impairment in the data dictionary

```
AD %>%
  filter(age >= 80 & cdrglob == 3.0)
```

```
## # A tibble: 6 × 57
##   id      diagnosis    age  educ female height weight bpsys
##   <chr>   <fct>      <dbl> <dbl> <fct>   <dbl>  <dbl> <dbl>
## 1 S544159 Dementia …    96    12 male     65.5    164   118
## 2 S738631 Dementia …    81    16 male     65.5    164   174
## 3 S863026 Dementia …    81    18 female   65.5    141   115
## 4 S689289 Dementia …    91    12 female   62      191   154
## 5 S278130 Dementia …    87    16 male     64      165   140
## 6 S219669 Dementia …    86    12 female   60       96   100
## # ℹ 46 more variables: naccgds <dbl>, delsev <dbl>, hallsev
## #   agitsev <dbl>, depdsev <dbl>, anxsev <dbl>, elatsev <dbl
```

```
AD %>%
  filter(age >= 80 & cdrglob == 3.0) %>%
  nrow()
```

```
## [1] 6
```

Here is when piping helps. We can pipe into other functions such as `nrow()`

Q. How many patients are diagnosed with questionable or mild dementia impairment (i.e. cdrglob > 0 and ≤ 1)?

```
AD %>%
  filter(cdrglob > 0 & cdrglob <= 1)
```

```
## # A tibble: 1,211 × 57
##    id      diagnosis   age  educ female height weight bpsys
##    <chr>   <fct>     <dbl> <dbl> <fct>   <dbl>  <dbl> <dbl>
##  1 S060833 Normal c…    74    12 male     65      233   148
##  2 S011143 Mild cog…    75    14 male     62      127   145
##  3 S122622 Dementia…    78    17 male     69      152   134
##  4 S297075 Dementia…    81    13 male     71      197   120
##  5 S194401 Mild cog…    75    16 female   68      110   145
##  6 S227329 Dementia…    75    12 male     66      180   128
##  7 S982276 Dementia…    56    16 male     67.5    166   130
##  8 S275920 Dementia…    81     9 female   64      184   150
```

Q. How many patients have moderate to severe impairment (cdrglob >= 2.0) and are female?

```
AD %>%
  filter(cdrglob >= 2 & female == "female") %>%
  nrow()
```

```
## [1] 34
```

We have done all sorts of selections, slicing, filtering on `AD` but it has not changed at all. Why do you think so?

```
glimpse(AD)
```

```
## Rows: 2,700
## Columns: 57
## $ id        <chr> "S060833", "S932623", "S755478", "S852291"
## $ diagnosis <fct> Normal cognition, Normal cognition, Normal
## $ age       <dbl> 74, 56, 77, 74, 75, 72, 64, 78, 73, 81, 66
## $ educ      <dbl> 12, 16, 18, 20, 14, 16, 16, 17, 18, 13, 16
## $ female    <fct> male, female, female, female, male, female
## $ height    <dbl> 65.0, 62.0, 65.0, 62.0, 62.0, 61.8, 60.0,
## $ weight    <dbl> 233, 110, 137, 112, 127, 141, 124, 152, 13
## $ bpsys     <dbl> 148, 110, 144, 120, 145, 107, 112, 134, 12
## $ bpdias    <dbl> 100, 75, 60, 60, 61, 65, 70, 74, 60, 70, 8
## $ hrate     <dbl> 72, 60, 64, 72, 58, 83, 76, 70, 60, 76, 60
```

Moving forward we are only going to use, `age`, `diagnosis` `female`, `bpsys` and `cdrglob`. Let's clean our data accordingly and move on with the smaller `AD` data that we need.

```
AD %>%
  select(age, diagnosis,
         female, bpsys,
         cdrglob)
```

```
## # A tibble: 2,700 × 5
##      age diagnosis                female bpsys cdrglob
##    <dbl> <fct>                    <fct>  <dbl>   <dbl>
##  1    74 Normal cognition         male     148     0.5
##  2    56 Normal cognition         female   110     0
##  3    77 Normal cognition         female   144     0
##  4    74 Normal cognition         female   120     0
##  5    75 Mild cognitive impairment male     145     0.5
```

Moving forward we are only going to use, `age`, `diagnosis` `female`, `bpsys` and `cdrglob`. Let's clean our data accordingly and move on with the smaller `AD` data that we need.

```
AD <-
  AD %>%
  select(age, diagnosis,
         female, bpsys,
         cdrglob)
```

```
glimpse(AD)
```

```
## Rows: 2,700
## Columns: 5
## $ age       <dbl> 74, 56, 77, 74, 75, 72, 64, 78, 73, 81, 66
## $ diagnosis <fct> Normal cognition, Normal cognition, Normal
## $ female    <fct> male, female, female, female, male, female
## $ bpsys     <dbl> 148, 110, 144, 120, 145, 107, 112, 134, 12
## $ cdrglob   <dbl> 0.5, 0.0, 0.0, 0.0, 0.5, 0.0, 0.0, 0.5, 0.
```

`mutate()` adds new variables and preserves existing ones

```
AD <-
  AD %>%
  mutate(age_days = 365*age)

colnames(AD)
```

```
## [1] "age"       "diagnosis" "female"    "bpsys"     "cdrglob
```

Grouping Data

Question:

> Do females have higher or lower CDRGLOB overall when compared with the males?

The function group_by() from dplyr groups the rows by the unique values in the column specified to it. Note that there is no perceptible change to the dataset after running group_by(), until another dplyr verb such as mutate(), summarise(), or arrange() is applied on the “grouped” data frame.

```
AD
```

```
## # A tibble: 2,700 × 6
##      age diagnosis                female bpsys cdrglob age_
##    <dbl> <fct>                    <fct>  <dbl>   <dbl>    <
##  1    74 Normal cognition         male     148     0.5    2
##  2    56 Normal cognition         female   110     0      2
##  3    77 Normal cognition         female   144     0      2
##  4    74 Normal cognition         female   120     0      2
##  5    75 Mild cognitive impairment male     145     0.5    2
##  6    72 Normal cognition         female   107     0      2
##  7    64 Normal cognition         female   112     0      2
```

Once we group the data, we won't see much difference other than `Groups: female [2]` statement, everything else will be similar.

```
AD %>%
  group_by(female)
```

```
## # A tibble: 2,700 × 6
## # Groups:   female [2]
##      age diagnosis                female bpsys cdrglob age_
##    <dbl> <fct>                    <fct>  <dbl>   <dbl>    <
##  1    74 Normal cognition         male     148     0.5    2
##  2    56 Normal cognition         female   110     0      2
##  3    77 Normal cognition         female   144     0      2
##  4    74 Normal cognition         female   120     0      2
##  5    75 Mild cognitive impairment male     145     0.5    2
##  6    72 Normal cognition         female   107     0      2
##  7    64 Normal cognition         female   112     0      2
```

```
AD %>%
  group_by(female) %>%
  summarize(median(cdrglob, na.rm = TRUE))
```

```
## # A tibble: 2 × 2
##   female `median(cdrglob, na.rm = TRUE)`
##   <fct>                            <dbl>
## 1 female                             0
## 2 male                               0.5
```

We can also calculate other descriptives as well as number of observations for each group.

```
AD %>%
  group_by(female) %>%
  summarize(med_cdrglob = median(cdrglob, na.rm = TRUE),
            mean_cdrglob = mean(cdrglob, na.rm = TRUE),
            n_cdrglob = n())
```

```
## # A tibble: 2 × 4
##   female med_cdrglob mean_cdrglob n_cdrglob
##   <fct>        <dbl>        <dbl>     <int>
## 1 female         0          0.271      1549
## 2 male           0.5        0.417      1151
```