데이터 탐색 & 전처리

Code ▾

** 이 문서는 R markdown으로 작성되었습니다.

데이터 불러오기

- 경로에 맞는 파일 이름 목록 가져오기

Hide

```
dir = "C:/data/health"

# 경로에서 패턴에 맞는 파일 이름 찾기
files <- list.files(path = dir, pattern = ".csv")
files
```

```
 [1] "2010.csv" "2011.csv" "2012.csv" "2013.csv"
 [5] "2014.csv" "2015.csv" "2016.csv" "2017.csv"
 [9] "2018.csv" "2019.csv" "2020.csv"
```

- 파일 이름 목록을 통해 데이터 불러오기

Hide

```
file <- paste(dir, "/", files, sep = "")
data_list <- lapply(file, read_csv)
```

```
Rows: 1000000 Columns: 34
── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (7): 음주여부, 구강검진 수검여부, 치아우식증유무, 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상, 치석
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당...
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 34
── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (7): 음주여부, 구강검진 수검여부, 치아우식증유무, 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상, 치석
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당...
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 34
── Column specification ──────────────────────────────────────────────────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (7): 음주여부, 구강검진 수검여부, 치아우식증유무, 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상, 치석
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당...
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
New names:
• `` -> `...35`
• `` -> `...36`
```

```
Warning: One or more parsing issues, call `problems()` on your data frame for details, e.g.:
  dat <- vroom(...)
  problems(dat)
```

```
Rows: 1000000 Columns: 36
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (7): 음주여부, 구강검진 수검여부, 치아우식증유무, 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상, 치석
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당...
lgl   (2): ...35, ...36
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 31
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (4): 음주여부, 구강검진 수검여부, 치아우식증유무, 치석유무
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당...
date  (1): 데이터 기준일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 31
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (4): 음주여부, 구강검진 수검여부, 치아우식증유무, 치석유무
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당...
date  (1): 데이터 기준일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 34
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (2): 구강검진 수검여부, 치석
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg단위), 허리둘레, 시력(좌),
시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당(...
lgl   (5): 음주여부, 치아우식증유무, 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 34
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (3): 음주여부, 구강검진수검여부, 치석
dbl  (26): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg단위), 허리둘레, 시력(좌),
시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당(...
lgl   (4): 치아우식증유무, 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 1000000 Columns: 34
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr   (3): 음주여부, 구강검진수검여부, 치석
dbl  (27): 기준년도, 가입자일련번호, 성별코드, 연령대코드(5세단위), 시도코드, 신장(5Cm단위), 체중(5Kg단위), 허리둘레, 시력(좌),
시력(우), 청력(좌), 청력(우), 수축기혈압, 이완기혈압, 식전혈당(...
lgl   (3): 결손치유무, 치아마모증유무, 제3대구치(사랑니)이상
date  (1): 데이터공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
Warning: One or more parsing issues, call `problems()` on your data frame for details, e.g.:
  dat <- vroom(...)
  problems(dat)
```

```
Rows: 1063619 Columns: 34
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr  (3): 결손치 유무, 치아마모증유무, 제3대구치(사랑니) 이상
dbl (30): 기준년도, 가입자 일련번호, 시도코드, 성별코드, 연령대 코드(5세단위), 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기 혈압, 이완기 혈압, ...
date (1): 데이터 공개일자

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 200118 Columns: 31
── Column specification ────────────────────────────────────────────────────────
────────────────────────────────────────────────────────────────────────────────
Delimiter: ","
chr  (1): 데이터 공개일자
dbl (30): 기준년도, 가입자 일련번호, 시도코드, 성별코드, 연령대 코드(5세단위), 신장(5Cm단위), 체중(5Kg 단위), 허리둘레, 시력
(좌), 시력(우), 청력(좌), 청력(우), 수축기 혈압, 이완기 혈압, 식...

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

데이터가 작성된 연도마다 속성, 기준이 다르므로 전처리를 따로 진행

2010 ~ 2013년 데이터 전처리

- 데이터 확인

Hide

```
data_2010 <- data.frame(data_list[1])
head(data_2010)
```

	기준년도 <dbl>	가입자일련번호 <dbl>	성별코드 <dbl>	연령대코드.5세단위. <dbl>	시도코드 <dbl>	신장.5Cm단위. <dbl>	체중.5Kg.단위. <dbl>	허리둘레 <dbl>	시력.좌. <dbl>
1	2010	141216	2	14	45	125	25	56	1.2
2	2010	393051	1	14	45	125	25	54	1.0
3	2010	78703	1	14	29	125	30	71	1.0
4	2010	7861	1	14	27	130	25	59	0.8
5	2010	5978	1	14	45	130	25	58	0.7
6	2010	8094	1	14	11	130	25	58	0.6

6 rows | 1-10 of 34 columns

Hide

```
data_2011 <- data.frame(data_list[2])
head(data_2011)
```

	기준년도 <dbl>	가입자일련번호 <dbl>	성별코드 <dbl>	연령대코드.5세단위. <dbl>	시도코드 <dbl>	신장.5Cm단위. <dbl>	체중.5Kg.단위. <dbl>	허리둘레 <dbl>	시력.좌. <dbl>
1	2011	762544	2	1	47	145	40	66	1.2
2	2011	56745	2	1	26	145	45	68	0.9
3	2011	171067	2	1	26	145	45	65	1.0
4	2011	196496	2	1	27	145	45	68	0.9
5	2011	812525	2	1	46	150	35	56	0.7
6	2011	270309	2	1	26	150	35	58	1.0

6 rows | 1-10 of 34 columns

Hide

```
data_2012 <- data.frame(data_list[3])
head(data_2012)
```

	기준년도 <dbl>	가입자일련번호 <dbl>	성별코드 <dbl>	연령대코드.5세단위. <dbl>	시도코드 <dbl>	신장.5Cm단위. <dbl>	체중.5Kg.단위. <dbl>	허리둘레 <dbl>	시력.좌. <dbl>
1	2012	220721	2	14	46	150	45	72	0.2
2	2012	830677	2	14	45	145	30	68	0.1
3	2012	978884	1	1	26	140	40	68	0.4
4	2012	918718	2	1	11	140	40	66	0.7
5	2012	922362	2	1	28	140	40	77	0.6
6	2012	9910	2	1	41	140	45	69	0.4

6 rows | 1-10 of 34 columns

Hide

```
data_2013 <- data.frame(data_list[4])
head(data_2013)
```

	기준년도 <dbl>	가입자일련번호 <dbl>	성별코드 <dbl>	연령대코드.5세단위. <dbl>	시도코드 <dbl>	신장.5Cm단위. <dbl>	체중.5Kg.단위. <dbl>	허리둘레 <dbl>	시력.좌. <dbl>
1	2013	24193	2	14	11	145	35	83	0.1
2	2013	134496	1	1	27	140	35	68	0.8
3	2013	978782	2	1	29	145	35	68	9.9
4	2013	10313	2	1	43	145	40	62	1.0
5	2013	915645	2	1	41	145	40	64	0.8
6	2013	3494	2	1	11	145	45	72	0.8

6 rows | 1-10 of 36 columns

- 열 이름 변경

전처리 편의, 병합을 위해 열 이름을 변경한다.

Hide

```
colName <- c("연도", "번호", "성별", "연령대", "지역", "신장", "체중", "허리", "좌시력", "우시력", "좌청력", "우청력", "수축혈압", "이완혈압", "공복혈당", "총콜레스테롤", "트리글리세라이드", "HDL콜레스테롤", "LDL콜레스테롤", "혈색소", "요단백", "혈청크레아티닌", "AST", "ALT", "감마지티피", "흡연", "음주", "구강검진", "치아우식증", "결손치", "치아마모증", "제3대구치", "치석", "공개일자")
names(data_2010) <- colName
names(data_2011) <- colName
names(data_2012) <- colName
names(data_2013) <- colName
```

- 데이터 변경

2002 ~ 2013년의 데이터는 2014년 이후의 데이터와 연령대 코드가 다르다. 2014년 이후의 데이터 기준에 맞춰 2010 ~ 2013년의 데이터를 변경해야 한다.

Hide

```
for (i in 1:14){
  data_2010[data_2010$연령대 == i, "연령대"] = i + 4
  data_2011[data_2011$연령대 == i, "연령대"] = i + 4
  data_2012[data_2012$연령대 == i, "연령대"] = i + 4
  data_2013[data_2013$연령대 == i, "연령대"] = i + 4
}
head(data_2010)
```

	연도 <dbl>	번호 <dbl>	성별 <dbl>	연령대 <dbl>	지역 <dbl>	신장 <dbl>	체중 <dbl>	허리 <dbl>	좌시력 <dbl>
1	2010	141216	2	18	45	125	25	56	1.2
2	2010	393051	1	18	45	125	25	54	1.0
3	2010	78703	1	18	29	125	30	71	1.0
4	2010	7861	1	18	27	130	25	59	0.8
5	2010	5978	1	18	45	130	25	58	0.7
6	2010	8094	1	18	11	130	25	58	0.6

6 rows | 1-10 of 34 columns

결측치 처리

- 결측치 개수 확인(전처리 전)

Hide

```
cat("2010년 데이터 결측치 개수 : ", sum(is.na(data_2010)), "개", "₩n")
```

```
2010년 데이터 결측치 개수 :  3440909 개
```

Hide

```
cat("2011년 데이터 결측치 개수 : ", sum(is.na(data_2011)), "개", "₩n")
```

```
2011년 데이터 결측치 개수 :  3472684 개
```

Hide

```
cat("2012년 데이터 결측치 개수 : ", sum(is.na(data_2012)), "개", "₩n")
```

```
2012년 데이터 결측치 개수 :  3092383 개
```

Hide

```
cat("2013년 데이터 결측치 개수 : ", sum(is.na(data_2013)), "개", "₩n")
```

```
2013년 데이터 결측치 개수 :  5091607 개
```

대부분의 결측치는 구강검진을 하지 않은 경우 관련 항목이 결측치로 처리된 경우, 검진 항목이 다른 경우 등에서 발생했다. 건강검진 데이터의 특성상, 다른 데이터와 관련이 있을 확률이 있으므로 다른 값으로 대체할 수 없다. 따라서, "-" 기호를 사용해 결측치를 대체하였다.

Hide

```
data_2010[is.na(data_2010)] <- "-"
data_2011[is.na(data_2011)] <- "-"
data_2012[is.na(data_2012)] <- "-"
data_2013[is.na(data_2013)] <- "-"
```

- 결측치 개수 확인(전처리 후)

Hide

```
cat("2010년 데이터 결측치 개수 : ", sum(is.na(data_2010)), "개", "₩n")
```

```
2010년 데이터 결측치 개수 :  0 개
```

Hide

```
cat("2011년 데이터 결측치 개수 : ", sum(is.na(data_2011)), "개", "₩n")
```

```
2011년 데이터 결측치 개수 :  0 개
```

Hide

```
cat("2012년 데이터 결측치 개수 : ", sum(is.na(data_2012)), "개", "₩n")
```

```
2012년 데이터 결측치 개수 :  0 개
```

Hide

```
cat("2013년 데이터 결측치 개수 : ", sum(is.na(data_2013)), "개", "₩n")
```

```
2013년 데이터 결측치 개수 :  0 개
```