

# vzeinali(vzeinali)\_Midterm\_QBS181

Vala Zeinali

10/30/2020

## Solutions

### Loading in Data

Let's start by loading in our data set. We will use sasxport to load in the df.

```
library(Hmisc)
```

```
## Warning: package 'Hmisc' was built under R version 3.6.2
## Loading required package: lattice
## Loading required package: survival
## Warning: package 'survival' was built under R version 3.6.2
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##      format.pval, units
```

```
library(sqldf)
```

```
## Loading required package: gsubfn
## Loading required package: proto
## Loading required package: RSQLite
## Warning: package 'RSQLite' was built under R version 3.6.2
df <- sasxport.get("/Users/valazeinali/Desktop/Dartmouth/Fall 2020/Data Wrangle/DIQ_I.XPT")
## Processing SAS dataset DIQ_I      ..
```

### Summary of raw data

To get a feel for what we are working with

```
summary(df)
```

```
##      seqn      diq010      did040      diq160
##  Min.   :83732  Min.   :1.000  Min.   :  2.00  Min.   :1.000
##  1st Qu.:86238  1st Qu.:2.000  1st Qu.: 40.00  1st Qu.:2.000
```

##	Median :88732	Median :2.000	Median : 50.00	Median :2.000	
##	Mean :88727	Mean :1.929	Mean : 62.56	Mean :1.928	
##	3rd Qu.:91218	3rd Qu.:2.000	3rd Qu.: 60.00	3rd Qu.:2.000	
##	Max. :93702	Max. :9.000	Max. :999.00	Max. :9.000	
##			NA's :8722	NA's :3530	
##	diq170	diq172	diq175a	diq175b	diq175c
##	Min. :1.000	Min. :1.000	Min. :10	Min. :11	Min. :12
##	1st Qu.:2.000	1st Qu.:1.000	1st Qu.:10	1st Qu.:11	1st Qu.:12
##	Median :2.000	Median :2.000	Median :10	Median :11	Median :12
##	Mean :1.876	Mean :1.843	Mean :10	Mean :11	Mean :12
##	3rd Qu.:2.000	3rd Qu.:2.000	3rd Qu.:10	3rd Qu.:11	3rd Qu.:12
##	Max. :9.000	Max. :9.000	Max. :10	Max. :11	Max. :12
##	NA's :3389	NA's :3389	NA's :8389	NA's :8974	NA's :9415
##	diq175d	diq175e	diq175f	diq175g	diq175h
##	Min. :13	Min. :14	Min. :15	Min. :16	Min. :17
##	1st Qu.:13	1st Qu.:14	1st Qu.:15	1st Qu.:16	1st Qu.:17
##	Median :13	Median :14	Median :15	Median :16	Median :17
##	Mean :13	Mean :14	Mean :15	Mean :16	Mean :17
##	3rd Qu.:13	3rd Qu.:14	3rd Qu.:15	3rd Qu.:16	3rd Qu.:17
##	Max. :13	Max. :14	Max. :15	Max. :16	Max. :17
##	NA's :9116	NA's :9420	NA's :9522	NA's :9295	NA's :9317
##	diq175i	diq175j	diq175k	diq175l	diq175m
##	Min. :18	Min. :19	Min. :20	Min. :21	Min. :22
##	1st Qu.:18	1st Qu.:19	1st Qu.:20	1st Qu.:21	1st Qu.:22
##	Median :18	Median :19	Median :20	Median :21	Median :22
##	Mean :18	Mean :19	Mean :20	Mean :21	Mean :22
##	3rd Qu.:18	3rd Qu.:19	3rd Qu.:20	3rd Qu.:21	3rd Qu.:22
##	Max. :18	Max. :19	Max. :20	Max. :21	Max. :22
##	NA's :9498	NA's :9425	NA's :9544	NA's :9516	NA's :9437
##	diq175n	diq175o	diq175p	diq175q	diq175r
##	Min. :23	Min. :24	Min. :25	Min. :26	Min. :27
##	1st Qu.:23	1st Qu.:24	1st Qu.:25	1st Qu.:26	1st Qu.:27
##	Median :23	Median :24	Median :25	Median :26	Median :27
##	Mean :23	Mean :24	Mean :25	Mean :26	Mean :27
##	3rd Qu.:23	3rd Qu.:24	3rd Qu.:25	3rd Qu.:26	3rd Qu.:27
##	Max. :23	Max. :24	Max. :25	Max. :26	Max. :27
##	NA's :9466	NA's :9441	NA's :9458	NA's :9439	NA's :9552
##	diq175s	diq175t	diq175u	diq175v	diq175w
##	Min. :28	Min. :29	Min. :30	Min. :31	Min. :32
##	1st Qu.:28	1st Qu.:29	1st Qu.:30	1st Qu.:31	1st Qu.:32
##	Median :28	Median :29	Median :30	Median :31	Median :32
##	Mean :28	Mean :29	Mean :30	Mean :31	Mean :32
##	3rd Qu.:28	3rd Qu.:29	3rd Qu.:30	3rd Qu.:31	3rd Qu.:32
##	Max. :28	Max. :29	Max. :30	Max. :31	Max. :32
##	NA's :9544	NA's :9470	NA's :9496	NA's :9564	NA's :9571
##	diq175x	diq180	diq050	did060	diq060u
##	Min. :33	Min. :1.000	Min. :1.000	Min. : 1.0	Min. :1.00
##	1st Qu.:33	1st Qu.:1.000	1st Qu.:2.000	1st Qu.: 3.0	1st Qu.:2.00
##	Median :33	Median :2.000	Median :2.000	Median : 7.0	Median :2.00
##	Mean :33	Mean :1.748	Mean :1.975	Mean : 29.8	Mean :1.88
##	3rd Qu.:33	3rd Qu.:2.000	3rd Qu.:2.000	3rd Qu.: 15.0	3rd Qu.:2.00
##	Max. :33	Max. :9.000	Max. :9.000	Max. :999.0	Max. :2.00
##	NA's :9574	NA's :3389		NA's :9319	NA's :9326
##	diq070	diq230	diq240	did250	

```
## Min. :1.000 Min. :1.000 Min. :1.000 Min. : 0.00
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 2.00
## Median :2.000 Median :4.000 Median :1.000 Median : 4.00
## Mean :1.583 Mean :3.236 Mean :1.246 Mean : 51.14
## 3rd Qu.:2.000 3rd Qu.:5.000 3rd Qu.:1.000 3rd Qu.: 4.00
## Max. :9.000 Max. :9.000 Max. :2.000 Max. :9999.00
## NA's :8060 NA's :8722 NA's :8722 NA's :8932
## did260 diq260u diq275 diq280
## Min. : 0.000 Min. :1.000 Min. :1.000 Min. : 2.0
## 1st Qu.: 1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.: 6.8
## Median : 1.000 Median :1.000 Median :1.000 Median : 8.1
## Mean : 1.657 Mean :1.578 Mean :1.708 Mean :373.3
## 3rd Qu.: 2.000 3rd Qu.:2.000 3rd Qu.:1.000 3rd Qu.:999.0
## Max. :15.000 Max. :4.000 Max. :9.000 Max. :999.0
## NA's :8726 NA's :8914 NA's :8722 NA's :8934
## diq291 diq300s diq300d did310s did310d
## Min. : 1.0 Min. : 80 Min. : 17 Min. : 80 Min. : 18
## 1st Qu.: 1.0 1st Qu.:124 1st Qu.: 75 1st Qu.:120 1st Qu.: 80
## Median : 2.0 Median :140 Median : 86 Median :6666 Median :6666
## Mean :19.2 Mean :3702 Mean :3947 Mean :4858 Mean :5086
## 3rd Qu.: 6.0 3rd Qu.:9999 3rd Qu.:9999 3rd Qu.:6666 3rd Qu.:9999
## Max. :99.0 Max. :9999 Max. :9999 Max. :9999 Max. :9999
## NA's :8934 NA's :8730 NA's :8730 NA's :8729 NA's :8729
## did320 did330 did341 did350
## Min. : 4 Min. : 6 Min. : 0.0 Min. : 0.00
## 1st Qu.:7777 1st Qu.:6666 1st Qu.: 0.0 1st Qu.: 1.00
## Median :9999 Median :9999 Median : 2.0 Median : 1.00
## Mean :8200 Mean :7777 Mean :127.6 Mean : 60.52
## 3rd Qu.:9999 3rd Qu.:9999 3rd Qu.: 4.0 3rd Qu.: 1.00
## Max. :9999 Max. :9999 Max. :9999.0 Max. :9999.00
## NA's :8729 NA's :8822 NA's :8732 NA's :8732
## diq350u diq360 diq080
## Min. :1.000 Min. :1.000 Min. :1.000
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:2.000
## Median :1.000 Median :2.000 Median :2.000
## Mean :1.409 Mean :2.681 Mean :1.837
## 3rd Qu.:2.000 3rd Qu.:3.000 3rd Qu.:2.000
## Max. :4.000 Max. :9.000 Max. :9.000
## NA's :8876 NA's :8729 NA's :8729
```

## Datatype Conversion

From our summary we notice that there are descriptive statistics on these columns. Unfortunately, we want these columns to be character type, for they are labels and are categorical. Some variables are quantitative in nature, however, are actually modeled qualitatively in this study.

```
temp<- supply(df, as.character) # making data qualitative "We dont wants integers"
df <- data.frame(temp)
head(df)
```

```
##      seqn diq010 did040 diq160 diq170 diq172 diq175a diq175b diq175c diq175d
## 1 83732      1     46  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 2 83733      2  <NA>      2      2      2  <NA>  <NA>  <NA>  <NA>
## 3 83734      1     52  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 4 83735      2  <NA>      1      1      2  <NA>  <NA>  <NA>  <NA>
```

```
## 5 83736      2 <NA>      2      2      1 <NA> <NA> <NA>      13
## 6 83737      2 <NA>      2      2      2 <NA> <NA> <NA> <NA>
##      diq175e diq175f diq175g diq175h diq175i diq175j diq175k diq175l diq175m
## 1      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 4      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 5      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      21      <NA>
## 6      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
##      diq175n diq175o diq175p diq175q diq175r diq175s diq175t diq175u diq175v
## 1      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 4      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 5      <NA>      24      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 6      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
##      diq175w diq175x diq180 diq050 did060 diq060u diq070 diq230 diq240 did250
## 1      <NA>      <NA>      <NA>      1      10      2      1      4      1      2
## 2      <NA>      <NA>      2      2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3      <NA>      <NA>      <NA>      2      <NA>      <NA>      2      3      2      <NA>
## 4      <NA>      <NA>      1      2      <NA>      <NA>      2      <NA>      <NA>      <NA>
## 5      <NA>      <NA>      2      2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 6      <NA>      <NA>      1      2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
##      did260 diq260u diq275 diq280 diq291 diq300s diq300d did310s did310d did320
## 1      1      1      1      7.3      2      134      72      6666      6666      9999
## 2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3      2      2      1      999      2      141      68      140      60      9999
## 4      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 5      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 6      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
##      did330 did341 did350 diq350u diq360 diq080
## 1      6666      0      1      1      2      2
## 2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3      9999      2      2      2      2      2
## 4      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 5      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 6      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
```

## Data Selection with SQL

I want to drop columns that have 99.5% missing values

```
print("A column will be dropped if it has less than 30 values")
```

```
## [1] "A column will be dropped if it has less than 30 values"
```

```
print(length(df$seqn) - length(df$seqn)*.996866841) # I want atleast 30 values in each column to be abl
```

```
## [1] 30
```

```
names <- data.frame(sapply(df, function(x) sum(is.na(x))/length(df$seqn)))
print(names > .996866841) #true columns will be dropped in SQL (next block)
```

```
##      sapply.df..function.x..sum.is.na.x...length.df.seqn..
## seqn                                                         FALSE
## diq010                                                        FALSE
```

## did040	FALSE
## diq160	FALSE
## diq170	FALSE
## diq172	FALSE
## diq175a	FALSE
## diq175b	FALSE
## diq175c	FALSE
## diq175d	FALSE
## diq175e	FALSE
## diq175f	FALSE
## diq175g	FALSE
## diq175h	FALSE
## diq175i	FALSE
## diq175j	FALSE
## diq175k	FALSE
## diq175l	FALSE
## diq175m	FALSE
## diq175n	FALSE
## diq175o	FALSE
## diq175p	FALSE
## diq175q	FALSE
## diq175r	TRUE
## diq175s	FALSE
## diq175t	FALSE
## diq175u	FALSE
## diq175v	TRUE
## diq175w	TRUE
## diq175x	TRUE
## diq180	FALSE
## diq050	FALSE
## did060	FALSE
## diq060u	FALSE
## diq070	FALSE
## diq230	FALSE
## diq240	FALSE
## did250	FALSE
## did260	FALSE
## diq260u	FALSE
## diq275	FALSE
## diq280	FALSE
## diq291	FALSE
## diq300s	FALSE
## diq300d	FALSE
## did310s	FALSE
## did310d	FALSE
## did320	FALSE
## did330	FALSE
## did341	FALSE
## did350	FALSE
## diq350u	FALSE
## diq360	FALSE
## diq080	FALSE

Now we want to get rid of the unwanted columns.

```
df1 <- sqldf("SELECT * FROM df ") #making copy of our data to preform data cleaning below
df = subset(df1, select = -c(di175r, di175v, di175w, di175x)) #dropping unwanted columns mathematica
head(df)
```

```
##      seqn diq010 did040 diq160 diq170 diq172 diq175a diq175b diq175c diq175d
## 1 83732      1      46   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 2 83733      2   <NA>      2      2      2   <NA>   <NA>   <NA>   <NA>
## 3 83734      1      52   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 4 83735      2   <NA>      1      1      2   <NA>   <NA>   <NA>   <NA>
## 5 83736      2   <NA>      2      2      1   <NA>   <NA>   <NA>      13
## 6 83737      2   <NA>      2      2      2   <NA>   <NA>   <NA>   <NA>
##      diq175e diq175f diq175g diq175h diq175i diq175j diq175k diq175l diq175m
## 1   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 3   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 4   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 5   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>      21   <NA>
## 6   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
##      diq175n diq175o diq175p diq175q diq175s diq175t diq175u diq180 diq050 did060
## 1   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>      1      10
## 2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>      2      2   <NA>
## 3   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>      2   <NA>
## 4   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>      1      2   <NA>
## 5   <NA>      24   <NA>   <NA>   <NA>   <NA>   <NA>      2      2   <NA>
## 6   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>      1      2   <NA>
##      diq060u diq070 diq230 diq240 did250 did260 diq260u diq275 diq280 diq291
## 1          2          1          4          1          2          1          1          1      7.3      2
## 2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 3   <NA>      2          3          2   <NA>      2          2          1     999      2
## 4   <NA>      2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 5   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 6   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
##      diq300s diq300d did310s did310d did320 did330 did341 did350 diq350u diq360
## 1       134       72     6666     6666    9999    6666         0         1         1         2
## 2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 3       141       68      140        60    9999    9999         2         2         2         2
## 4   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 5   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 6   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
##      diq080
## 1          2
## 2   <NA>
## 3          2
## 4   <NA>
## 5   <NA>
## 6   <NA>
```

## OneHotEncoding for future ML purposes

I am making two data sets, one for columns that represent unique questions and the other represents one question with multiple responses. The “factors” table below is set up for feature selection for future ML purposes.

```
factors <- sqldf("SELECT
seqn,
diq175a,
diq175b,
diq175c,
diq175d,
diq175e,
diq175f,
diq175g,
diq175h,
diq175i,
diq175j,
diq175k,
diq175l,
diq175m,
diq175n,
diq175o,
diq175p,
diq175q,
diq175s,
diq175t,
diq175u FROM df ")
```

```
head(factors)
```

```
##      seqn diq175a diq175b diq175c diq175d diq175e diq175f diq175g diq175h diq175i
## 1 83732      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 2 83733      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3 83734      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 4 83735      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 5 83736      <NA>      <NA>      <NA>      13      <NA>      <NA>      <NA>      <NA>      <NA>
## 6 83737      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
##      diq175j diq175k diq175l diq175m diq175n diq175o diq175p diq175q diq175s
## 1      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 2      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 3      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 4      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
## 5      <NA>      <NA>      21      <NA>      <NA>      24      <NA>      <NA>      <NA>
## 6      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>      <NA>
##      diq175t diq175u
## 1      <NA>      <NA>
## 2      <NA>      <NA>
## 3      <NA>      <NA>
## 4      <NA>      <NA>
## 5      <NA>      <NA>
## 6      <NA>      <NA>
```

```
library(dbplyr)
```

```
## Warning: package 'dbplyr' was built under R version 3.6.2
```

```
#replace all N/A's with zeros
```

```
i <- sapply(factors, is.factor) # Identify all factor variables in your data
factors[i] <- lapply(factors[i], as.character) # Convert factors to character variables
factors[is.na(factors)] <- 0 # Replace NA with 0, as shown in Example 1
```

```
factors[i] <- lapply(factors[i], as.factor) # Convert character columns back to factors
```

```
#one-hot-encoding
```

```
factors <- factors %>% dplyr::mutate_if(is.character, ~1 * (. != 0)) #replace all values that are not 0
```

```
factors$seqn <- df$seqn #need to restore the seqn data
```

```
head(factors)
```

```
##      seqn diq175a diq175b diq175c diq175d diq175e diq175f diq175g diq175h diq175i
## 1 83732      0      0      0      0      0      0      0      0      0
## 2 83733      0      0      0      0      0      0      0      0      0
## 3 83734      0      0      0      0      0      0      0      0      0
## 4 83735      0      0      0      0      0      0      0      0      0
## 5 83736      0      0      0      1      0      0      0      0      0
## 6 83737      0      0      0      0      0      0      0      0      0
##      diq175j diq175k diq175l diq175m diq175n diq175o diq175p diq175q diq175s
## 1      0      0      0      0      0      0      0      0      0
## 2      0      0      0      0      0      0      0      0      0
## 3      0      0      0      0      0      0      0      0      0
## 4      0      0      0      0      0      0      0      0      0
## 5      0      0      1      0      0      1      0      0      0
## 6      0      0      0      0      0      0      0      0      0
##      diq175t diq175u
## 1      0      0
## 2      0      0
## 3      0      0
## 4      0      0
## 5      0      0
## 6      0      0
```

```
#Summary
```

Now we have two tables, one table has columns that represent unique questions and the other represents one question with multiple responses. The “factors” table is set up for feature selection for future ML purposes.

```
head(factors)
```

```
##      seqn diq175a diq175b diq175c diq175d diq175e diq175f diq175g diq175h diq175i
## 1 83732      0      0      0      0      0      0      0      0      0
## 2 83733      0      0      0      0      0      0      0      0      0
## 3 83734      0      0      0      0      0      0      0      0      0
## 4 83735      0      0      0      0      0      0      0      0      0
## 5 83736      0      0      0      1      0      0      0      0      0
## 6 83737      0      0      0      0      0      0      0      0      0
##      diq175j diq175k diq175l diq175m diq175n diq175o diq175p diq175q diq175s
## 1      0      0      0      0      0      0      0      0      0
## 2      0      0      0      0      0      0      0      0      0
## 3      0      0      0      0      0      0      0      0      0
## 4      0      0      0      0      0      0      0      0      0
## 5      0      0      1      0      0      1      0      0      0
## 6      0      0      0      0      0      0      0      0      0
##      diq175t diq175u
## 1      0      0
## 2      0      0
## 3      0      0
## 4      0      0
## 5      0      0
```



```
## 6      0      0
```

```
head(df)
```

```
##      seqn diq010 did040 diq160 diq170 diq172 diq175a diq175b diq175c diq175d
## 1 83732      1     46  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 2 83733      2  <NA>      2      2      2  <NA>  <NA>  <NA>  <NA>
## 3 83734      1     52  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 4 83735      2  <NA>      1      1      2  <NA>  <NA>  <NA>  <NA>
## 5 83736      2  <NA>      2      2      1  <NA>  <NA>  <NA>     13
## 6 83737      2  <NA>      2      2      2  <NA>  <NA>  <NA>  <NA>
##      diq175e diq175f diq175g diq175h diq175i diq175j diq175k diq175l diq175m
## 1  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 2  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 3  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 4  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 5  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>     21  <NA>
## 6  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##      diq175n diq175o diq175p diq175q diq175s diq175t diq175u diq180 diq050 did060
## 1  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>      1     10
## 2  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>      2      2  <NA>
## 3  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>      2  <NA>
## 4  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>      1      2  <NA>
## 5  <NA>     24  <NA>  <NA>  <NA>  <NA>  <NA>      2      2  <NA>
## 6  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>      1      2  <NA>
##      diq060u diq070 diq230 diq240 did250 did260 diq260u diq275 diq280 diq291
## 1      2      1      4      1      2      1      1      1     7.3      2
## 2  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 3  <NA>      2      3      2  <NA>      2      2      1    999      2
## 4  <NA>      2  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 5  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 6  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##      diq300s diq300d did310s did310d did320 did330 did341 did350 diq350u diq360
## 1     134      72    6666    6666    9999    6666      0      1      1      2
## 2  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 3     141      68     140      60    9999    9999      2      2      2      2
## 4  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 5  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
## 6  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>
##      diq080
## 1      2
## 2  <NA>
## 3      2
## 4  <NA>
## 5  <NA>
## 6  <NA>
```