Análisis Base de Datos Simplificada

Tomás Sánchez Grigioni

13/9/2020

Análisis de Datos

Las variables con las que se va a trabajar son:

Nombre Variable Original	Nombre Variable Nuevo	Significado	Tipo Variable
healthCode	-	-	-
healthCode1	-	Es el identificador del paciente	-
ROW_ID	id	Es el identificador de fila	-
DIAGNOSTICO	diagnostico	Resultado del diagnóstico	Categórica
d1	d1	Mean of logarithmic F0 on a semitone frequency scale	Continua
d2	d2	Mean of the ratio of the energy of the spectral harmonic peak at the first formant's center frequency to the energy of the spectral peak at F0 in voiced regions	Continua
d3	d3	Coefficient of variation of the ratio of the energy of the spectral harmonic peak at the first formant's center frequency to the energy of the spectral peak at F0 in voiced regions	Continua
d4	d4	Mean of linear regression slope of the logarithmic power spectrum within 0–500 Hz band entropy.	Continua
d5	d5	Mean Jitter of the deviations in individual consecutive F0 period lengths.	Continua
medTimePoint	punto_medicacion	Punto de toma de medicación	Cualitativa
age	edad	Edad del paciente	Discreta
are-caretaker	cuidado	Esta con cuidado	Cualitativa
deep-brain-stimulation	estimulacion_cerebral	Presenta el tratamiento de estimulación cerebral	Cualitativa
diagnosis-year	anio_diagnostico	Año en que le diagnosticaron la enfermedad	Discreta
education	educacion	Nivel de educacion	Cualitativa
employment	trabajo	Trabajo del paciente	Cualitativa
gender	genero	Genero del paciente	Cualitativa
maritalStatus	estado_civil	Estado civil	Cualitativa

Nombre Variable Original	Nombre Variable Nuevo	Significado	Tipo Variable
medication-start-year	anio_medicacion	Año en que comenzo a medicarse	Discreta
onset-year	anio_enfermedad	Año en que se manifesto la enfermedad	Discreta
smartphone	facilidad_celular	Facilidad con la que usa el celular	Cualitativa
smoked	fumo	Si fumo	Cualitativa
surgery	cirugias	Si presenta cirugías	Cualitativa
years-smoking	anios_fumo	Cantidad de años que fumo	Discreta

Primero cargamos los datos, luego observamos como los trata R.

```
## tibble [2,218 x 38] (S3: tbl_df/tbl/data.frame)
##  $ id                                          : num [1:2218] 24665 21472 25155 22829 24848 ...
##  $ diagnostico                                 : num [1:2218] 1 0 1 1 1 1 1 1 1 0 ...
##  $ d1                                          : num [1:2218] 35.2 35.1 53.8 31.2 36.3 ...
##  $ d2                                          : num [1:2218] -179 -177 -179 -197 -118 ...
##  $ d3                                          : num [1:2218] 0.375 0.38 0.36 0.326 0.271 ...
##  $ d4                                          : num [1:2218] 0.0966 0.1159 0.1094 0.1032 0.1028 .
##  $ d5                                          : num [1:2218] 0.03211 0.02492 0.05817 0.06926 0.005
##  $ healthCode1                                 : chr [1:2218] "c9640e39-4b86-43d7-a3bb-b11838143f1
##  $ punto_medicacion                            : chr [1:2218] "Immediately before Parkinson medica
##  $ F0semitoneFrom27#5Hz_sma3nz_amean           : num [1:2218] 35.2 35.1 53.8 31.2 36.3 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_stddevNorm      : num [1:2218] 0.161 0.107 0.182 0.162 0.055 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_percentile20#0  : num [1:2218] 29.5 32.4 40.4 28 36 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_percentile50#0  : num [1:2218] 38.6 35.8 60.4 28.5 36 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_percentile80#0  : num [1:2218] 38.8 37.6 61.3 36.9 36.1 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_pctlrange0-2    : num [1:2218] 9.387 5.165 20.833 8.948 0.153 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_meanRisingSlope : num [1:2218] 114 118 429 419 241 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_stddevRisingSlope: num [1:2218] 56.1 107.3 689.4 321.4 241.5 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_meanFallingSlope : num [1:2218] 97.9 81 127.1 254.6 121.1 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_stddevFallingSlope: num [1:2218] 182.3 77.6 123 53.7 112.6 ...
##  $ jitterLocal_sma3nz_amean                    : num [1:2218] 0.03211 0.02492 0.05817 0.06926 0.005
##  $ jitterLocal_sma3nz_stddevNorm               : num [1:2218] 2.301 1.259 0.668 0.968 3.41 ...
##  $ shimmerLocaldB_sma3nz_amean                 : num [1:2218] 0.752 1.418 1.862 2.252 0.275 ...
##  $ shimmerLocaldB_sma3nz_stddevNorm            : num [1:2218] 1.076 0.564 0.269 0.587 2.213 ...
##  $ edad                                        : num [1:2218] 71 48 73 50 57 51 56 52 64 36 ...
##  $ cuidado                                     : chr [1:2218] "true" "false" "false" "false" ...
##  $ estimulacion_cerebral                       : chr [1:2218] "false" "NULL" "false" "false" ...
##  $ anio_diagnostico                            : chr [1:2218] "2011" "NULL" "2005" "2014" ...
##  $ educacion                                   : chr [1:2218] "High School Diploma/GED" "Some colle
##  $ trabajo                                     : chr [1:2218] "Retired" "Employment for wages" "Re
##  $ genero                                      : chr [1:2218] "Female" "Male" "Male" "Male" ...
##  $ estado_civil                                : chr [1:2218] "Married or domestic partnership" "Ma
##  $ anio_medicacion                             : chr [1:2218] "2011" "NULL" "2005" "0" ...
##  $ anio_enfermedad                             : chr [1:2218] "2011" "NULL" "2005" "2012" ...
##  $ facilidad_celular                           : chr [1:2218] "Neither easy nor difficult" "Very ea
##  $ fumo                                        : chr [1:2218] "false" "false" "true" "false" ...
##  $ cirugias                                    : chr [1:2218] "false" "NULL" "false" "false" ...
##  $ anios_fumo                                  : chr [1:2218] "NULL" "NULL" "4" "NULL" ...
##  $ healthCode                                  : chr [1:2218] "c9640e39-4b86-43d7-a3bb-b11838143f1
```

Los problemas que encontramos son:

- **diagnostico** no es un factor.
- **punto_medicacion** no es un factor.
- **cuidado** no es un factor.
- **estimulacion_cerebral** no es un factor.
- **anio_diagnostico** no es una fecha.
- **educacion** no es un factor.
- **trabajo** no es un factor.
- **genero** no es un factor.
- **estado_civil** no es un factor.
- **anio_medicacion** no es una fecha
- **anio_enfermedad** no es una fecha.
- **facilidad_celular** no es un factor.
- **fumo** no es un factor.
- **cirugias** no es un factor.
- **anios_fumo** no es un número.
- Además le cambiamos lo nombres a las variables para que sea más manejable.

```
## tibble [2,218 x 38] (S3: tbl_df/tbl/data.frame)
##  $ id                                          : num [1:2218] 24665 21472 25155 22829 24848 ...
##  $ diagnostico                                 : Factor w/ 2 levels "0","1": 2 1 2 2 2 2 2 2 2 1 .
##  $ d1                                          : num [1:2218] 35.2 35.1 53.8 31.2 36.3 ...
##  $ d2                                          : num [1:2218] -179 -177 -179 -197 -118 ...
##  $ d3                                          : num [1:2218] 0.375 0.38 0.36 0.326 0.271 ...
##  $ d4                                          : num [1:2218] 0.0966 0.1159 0.1094 0.1032 0.1028 .
##  $ d5                                          : num [1:2218] 0.03211 0.02492 0.05817 0.06926 0.005
##  $ healthCode1                                 : chr [1:2218] "c9640e39-4b86-43d7-a3bb-b11838143f1
##  $ punto_medicacion                            : Factor w/ 4 levels "Another time",..: 3 2 3 2 3 4
##  $ F0semitoneFrom27#5Hz_sma3nz_amean           : num [1:2218] 35.2 35.1 53.8 31.2 36.3 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_stddevNorm      : num [1:2218] 0.161 0.107 0.182 0.162 0.055 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_percentile20#0  : num [1:2218] 29.5 32.4 40.4 28 36 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_percentile50#0  : num [1:2218] 38.6 35.8 60.4 28.5 36 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_percentile80#0  : num [1:2218] 38.8 37.6 61.3 36.9 36.1 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_pctlrange0-2    : num [1:2218] 9.387 5.165 20.833 8.948 0.153 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_meanRisingSlope : num [1:2218] 114 118 429 419 241 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_stddevRisingSlope: num [1:2218] 56.1 107.3 689.4 321.4 241.5 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_meanFallingSlope : num [1:2218] 97.9 81 127.1 254.6 121.1 ...
##  $ F0semitoneFrom27#5Hz_sma3nz_stddevFallingSlope: num [1:2218] 182.3 77.6 123 53.7 112.6 ...
##  $ jitterLocal_sma3nz_amean                    : num [1:2218] 0.03211 0.02492 0.05817 0.06926 0.005
##  $ jitterLocal_sma3nz_stddevNorm               : num [1:2218] 2.301 1.259 0.668 0.968 3.41 ...
##  $ shimmerLocaldB_sma3nz_amean                 : num [1:2218] 0.752 1.418 1.862 2.252 0.275 ...
##  $ shimmerLocaldB_sma3nz_stddevNorm            : num [1:2218] 1.076 0.564 0.269 0.587 2.213 ...
##  $ edad                                        : num [1:2218] 71 48 73 50 57 51 56 52 64 36 ...
##  $ cuidado                                     : Factor w/ 2 levels "false","true": 2 1 1 1 1 1 1
##  $ estimulacion_cerebral                       : Factor w/ 2 levels "false","true": 1 NA 1 1 1 1 1
##  $ anio_diagnostico                            : num [1:2218] 2011 NA 2005 2014 2012 ...
##  $ educacion                                   : Factor w/ 8 levels "2-year college degree",..: 4 6
##  $ trabajo                                     : Factor w/ 7 levels "A homemaker",..: 5 3 5 6 6 3 5
##  $ genero                                      : Factor w/ 3 levels "Female","Male",..: 1 2 2 2 1 1
##  $ estado_civil                                : Factor w/ 6 levels "Divorced","Married or domesti
##  $ anio_medicacion                             : num [1:2218] 2011 NA 2005 0 2012 ...
##  $ anio_enfermedad                             : num [1:2218] 2011 NA 2005 2012 2011 ...
##  $ facilidad_celular                           : Factor w/ 5 levels "Difficult","Easy",..: 3 5 3 5
```

```
##  $ fumo                                          : Factor w/ 2 levels "false","true": 1 1 2 1 2 1 1 1
##  $ cirugias                                      : Factor w/ 2 levels "false","true": 1 NA 1 1 1 1 1 1
##  $ anios_fumo                                    : num [1:2218] NA NA 4 NA 15 NA NA 30 10 NA ...
##  $ healthCode                                    : chr [1:2218] "c9640e39-4b86-43d7-a3bb-b11838143f1
```

Ahora las variables se tratan correctamente. Calculamos las medidas resumen para las variables cuantitativas y creamos tablas de frecuencia para variables cualitativas

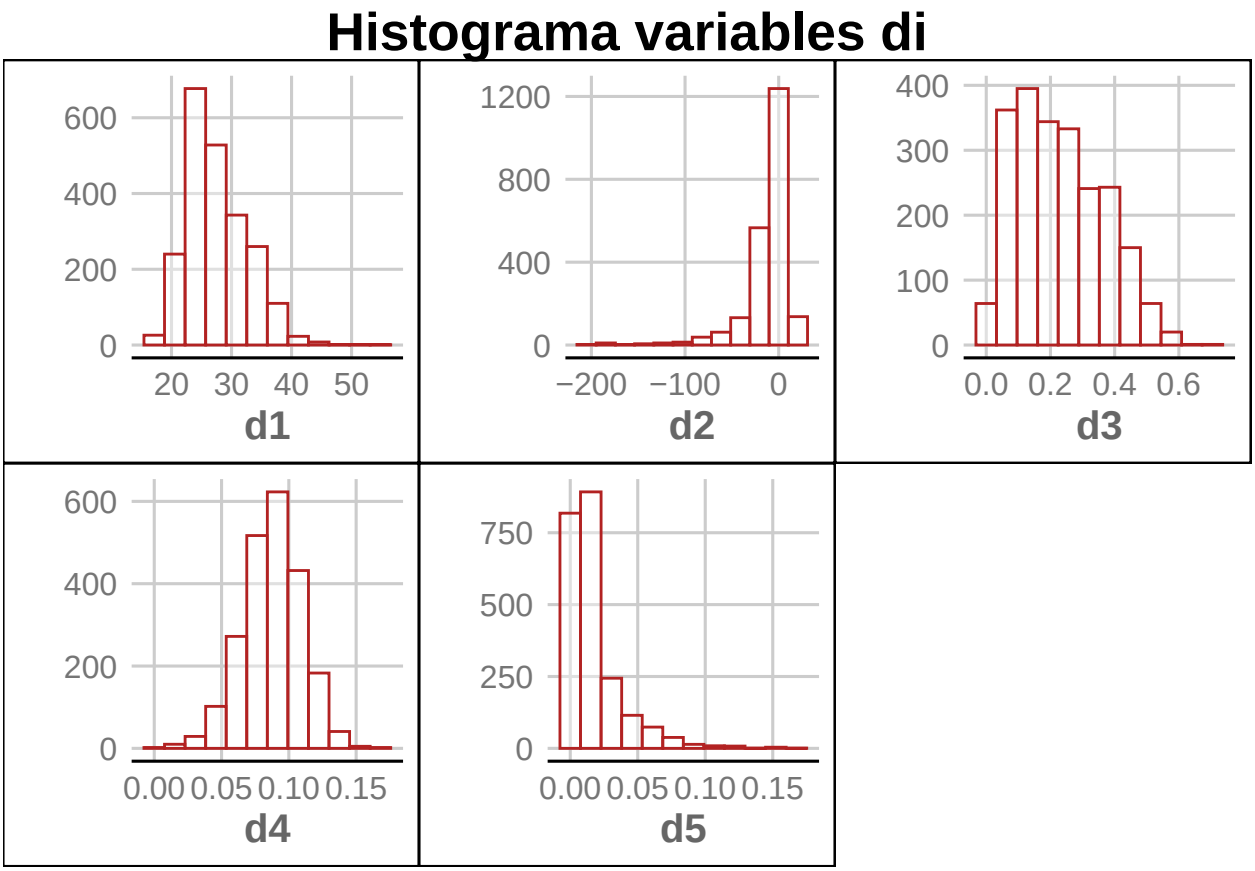
De las variables cuantitativas los problemas que se encuentran son que:

- **anio_medicacion** presenta una media en 1429.44 indicando que hay valores muy chicos que distorsionan la verdadera media. Luego, observamos que el primer cuartil tiene valor 0. Estos valores es posible que indiquen que esa persona no toma medicaciones y en su lugar se debería reemplazar estas observaciones con NA para que no interfieran.

Realizamos algunas verificaciones para determinar que no hay otros errores de este estilo.

Por la presencia de los 0s en la variable **anios_fumo** decidimos reemplazarlos por NAs. Además, completamos con NAs aquellos valores que sean anteriores a 1960 en la variable **anio_medicacion**.

Análisis Univariado

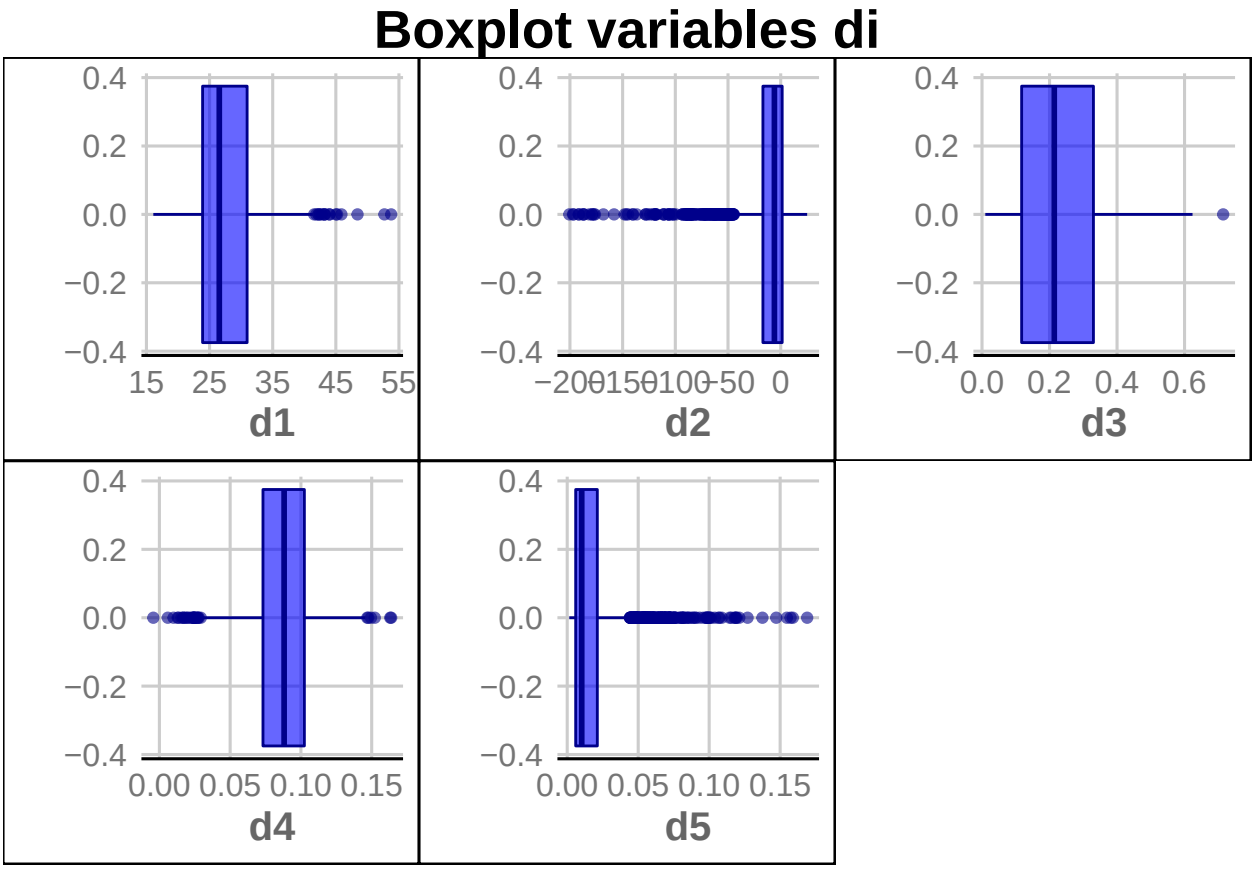


- En **d1** parece ser que los datos se concentran entre los valores de 20 y 40, teniendo algunas observaciones por fuera de este intervalo.

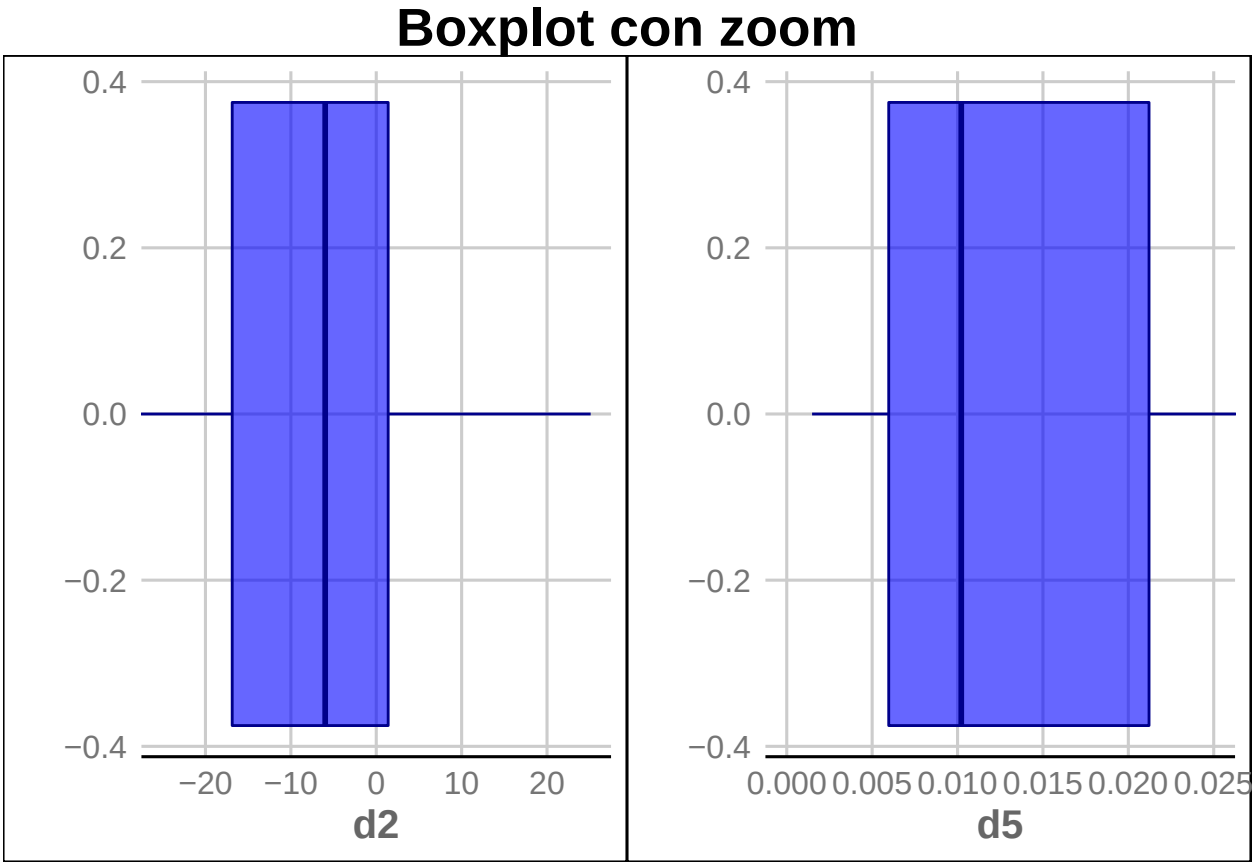
Table 2: Medidas Resumen de Variables Cuantitativas

Variable	Media	Mediana	Primer_Cua
d1	27.59	26.54	23
d2	-12.59	-5.98	-16
d3	0.23	0.21	0
d4	0.09	0.09	0
d5	0.02	0.01	0
F0semitoneFrom27#5Hz_sma3nz_amean	27.59	26.54	23
F0semitoneFrom27#5Hz_sma3nz_stddevNorm	0.07	0.05	0
F0semitoneFrom27#5Hz_sma3nz_percentile20#0	26.70	25.63	22
F0semitoneFrom27#5Hz_sma3nz_percentile50#0	27.44	26.40	23
F0semitoneFrom27#5Hz_sma3nz_percentile80#0	28.43	27.36	24
F0semitoneFrom27#5Hz_sma3nz_pctlrange0-2	1.74	0.46	0
F0semitoneFrom27#5Hz_sma3nz_meanRisingSlope	98.65	59.42	4
F0semitoneFrom27#5Hz_sma3nz_stddevRisingSlope	84.07	43.13	2
F0semitoneFrom27#5Hz_sma3nz_meanFallingSlope	57.49	26.12	4
F0semitoneFrom27#5Hz_sma3nz_stddevFallingSlope	57.40	24.08	2
jitterLocal_sma3nz_amean	0.02	0.01	0
jitterLocal_sma3nz_stddevNorm	1.94	1.79	1
shimmerLocaldB_sma3nz_amean	0.78	0.63	0
shimmerLocaldB_sma3nz_stddevNorm	0.90	0.82	0
edad	53.62	53.00	43
anio_diagnostico	2009.73	2011.00	2007
anio_medicacion	1429.44	2009.00	0
anio_enfermedad	2007.91	2010.00	2006
anios_fumo	14.32	12.00	5

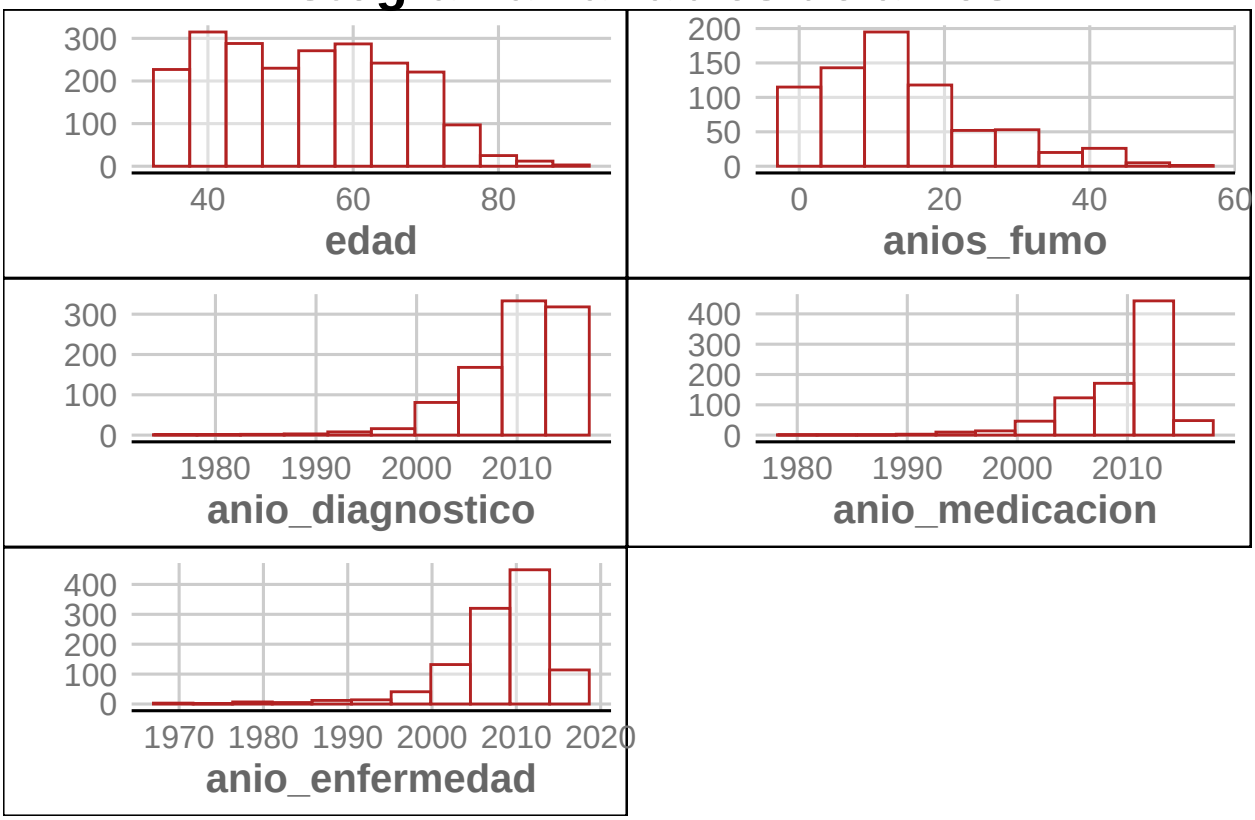
- En **d2** los datos parecen estar centrados en 0, pero presenta una cola con algunas observaciones para los números negativos.
- En **d3** todas las observaciones se encuentran distribuidas entre 0 y 0.6, con gran cantidad de observaciones en los valores intermedios.
- En **d4** todas las observaciones se encuentran distribuidas entre 0 y 0.15, pero con mayor cantidad de observaciones para los valores que se encuentran entre 0.05 0.10.
- En **d5** se presenta la mayor cantidad de observaciones en los valores más cercanos a 0 de esta variable. Teniendo también observaciones para valores mayores a 0.10, pero en menor medida.



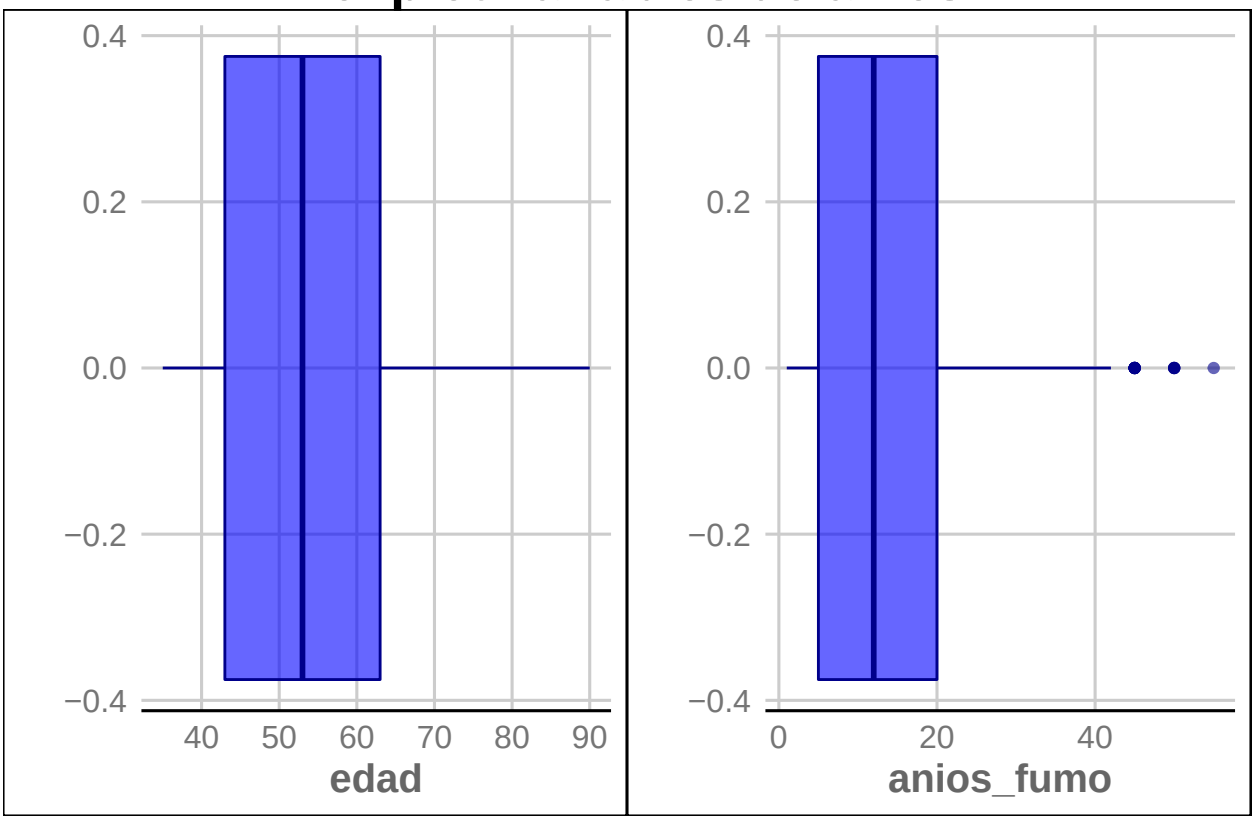
- En **d1** se tiene la mediana cercana a 25, y el 50% central de los datos parecen estar contenidos entre 25 y 30 aproximadamente. Hay gran presencia de valores extremos.
- En **d2** hay gran presencia de valores extremos por lo que se distorsiona el gráfico, posteriormente realizamos un zoom a la zona de interés.
- En **d3** la mediana pareciera estar cerca de 0.2 y el 50% central entre 0.1 y 0.3, aprozimadamente. Presenta un único valor extremo.
- En **d4** la mediana se encuentra cercana a 0.1, con el 50% central entre 0.08 y 0.1. Hay varios valores extremos.
- En **d5** hay gran presencia de valores extremos, por lo que realizamos un zoom para analizar.



- En **d2** la mediana se encuentra cercana a -5 y el 50% central de los datos entre -50 y 0, aproximadamente.
- En **d5** la mediana se encuentra aproximadamente en 0.010, con el 50% central entre 0.005 y 0.020 aproximadamente.

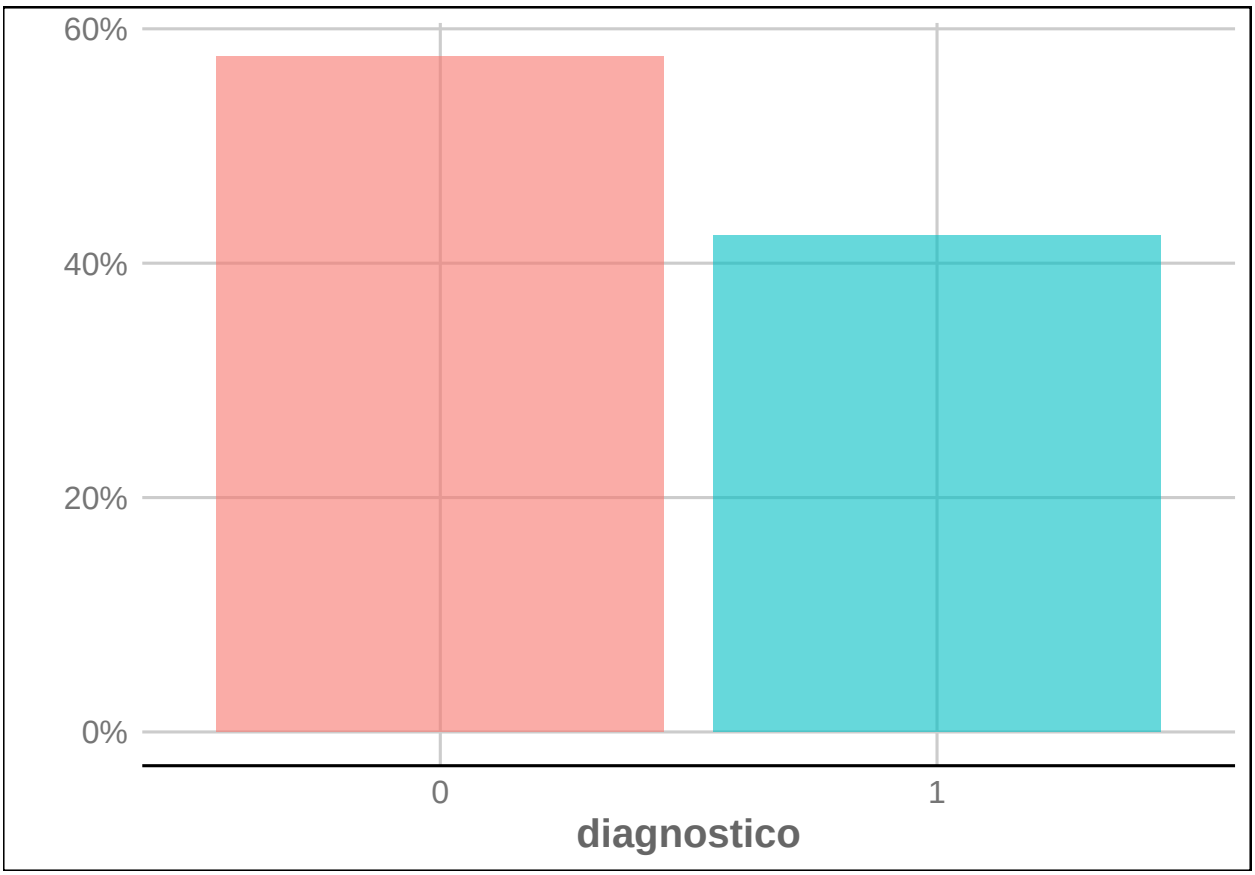


- En **edad** encontramos que las observaciones parecieran distribuirse uniformemente en el rango de 40 y 70, con algunas pocas observaciones extras para edades superiores a 70.
- En **anios_fumo** encontramos que existen observaciones de valores entre 1 y 50 prácticamente. Pero la mayo cantidad de observaciones se centran en los 15 años aproximadamente.
- En **anio_diagnostico** la mayor cantidad de observaciones se encuentran cerca del año 2010, pero existen algunas observaciones para los años menores a 2000.
- En **anio_medicacion** encontramos gran cantidad de observaciones cercanas al año 2010, y unas pocas para los años menores a 200.
- En **anio_enfermedad** la mayor cantidad de observaciones se encuentran cerca del año 2010, pero existen algunas observaciones para los años menores a 2000.



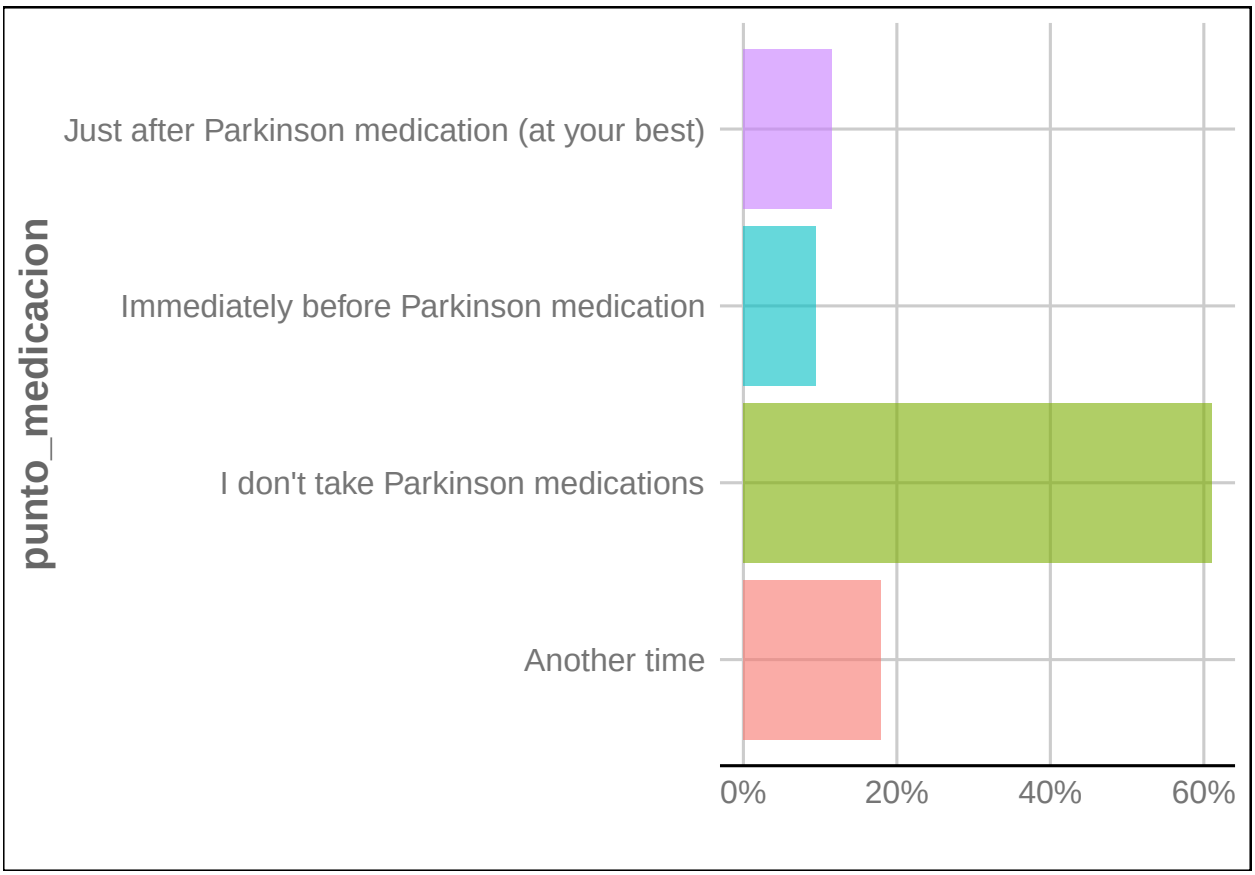
- En **edad** parece que la mediana se encuentra cercana a 52 años, con el 50% central de los datos desde los 45 hasta los 62 aproximadamente. No hay resencia de valores extremos.
- En **anios_fumo** la mediana parece encontrarse cerca de los 10 años. El 50% central de los datos va desde los 5 años hasta los 20. Existen algunos valores extremos.

diagnostico	Freq
0	1278
1	940



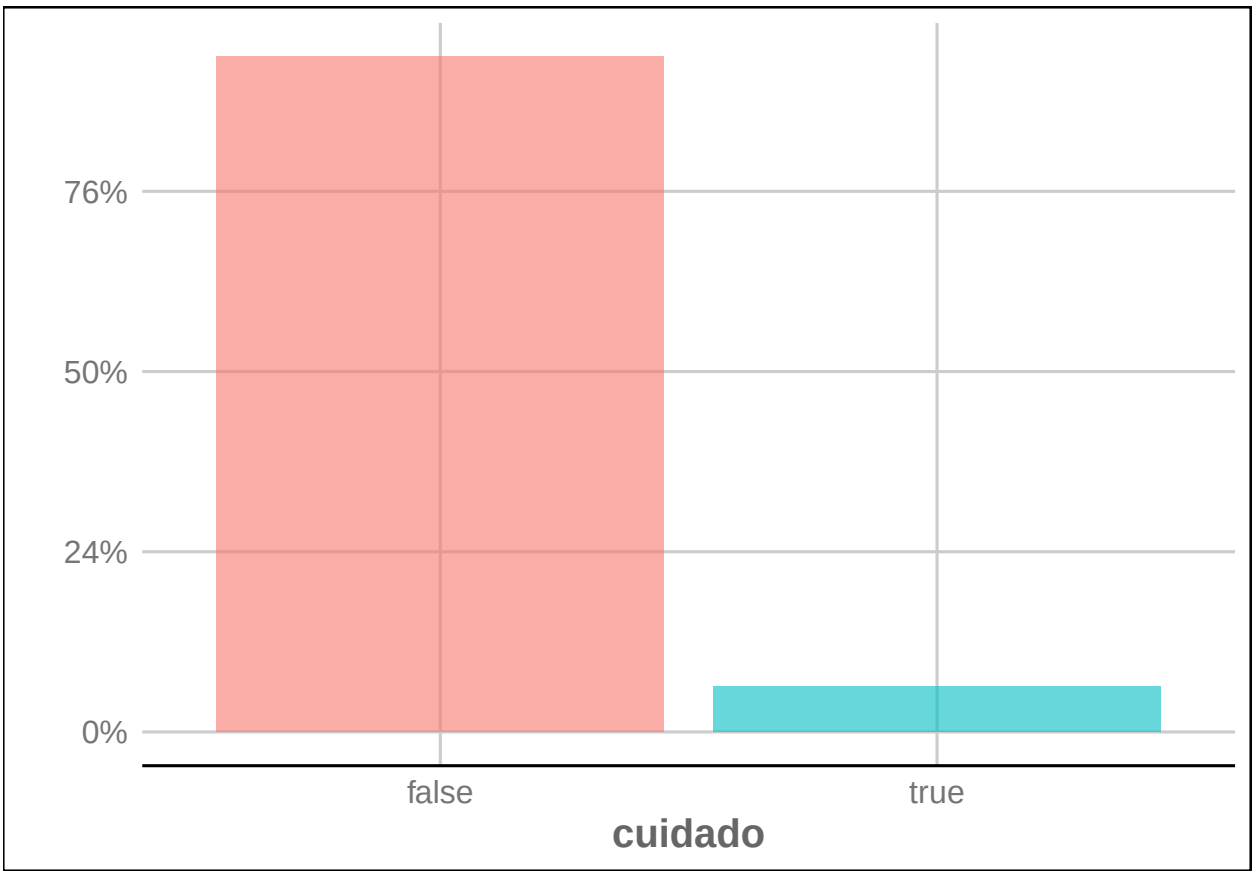
- En **diagnostico** más de la mitad de los datos corresponden a diagnósticos negativos, y el resto a diagnosticos positivos.

punto_medicacion	Freq
Another time	394
I don't take Parkinson medications	1343
Immediately before Parkinson medication	209
Just after Parkinson medication (at your best)	255



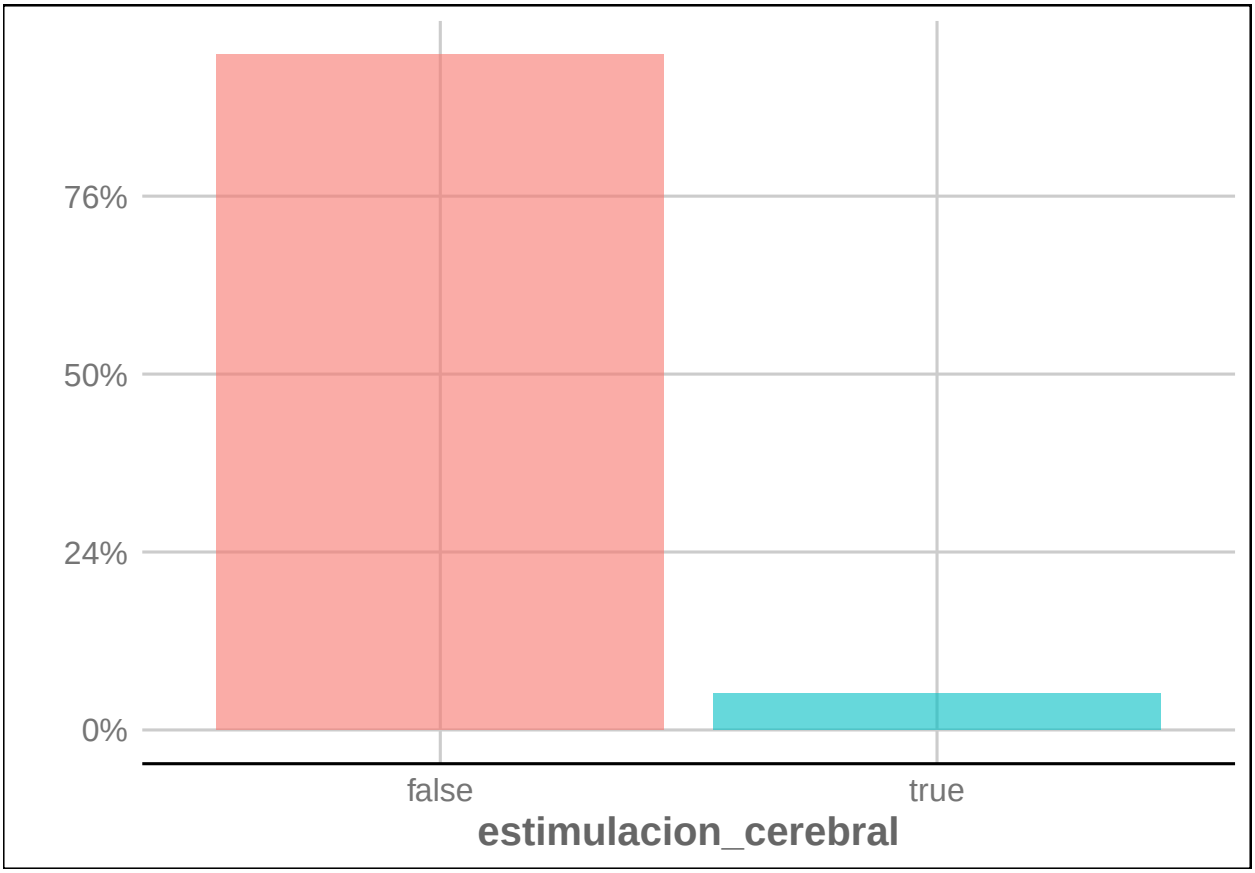
- En **punto_medicacion** el 60% de los datos corresponden a personas que no toman medicación para parkinson. Luego sigue, "Another time", "Just After Parkinson medication" y "Immediately before Parkinson medication".

cuidado	Freq
false	2075
true	140



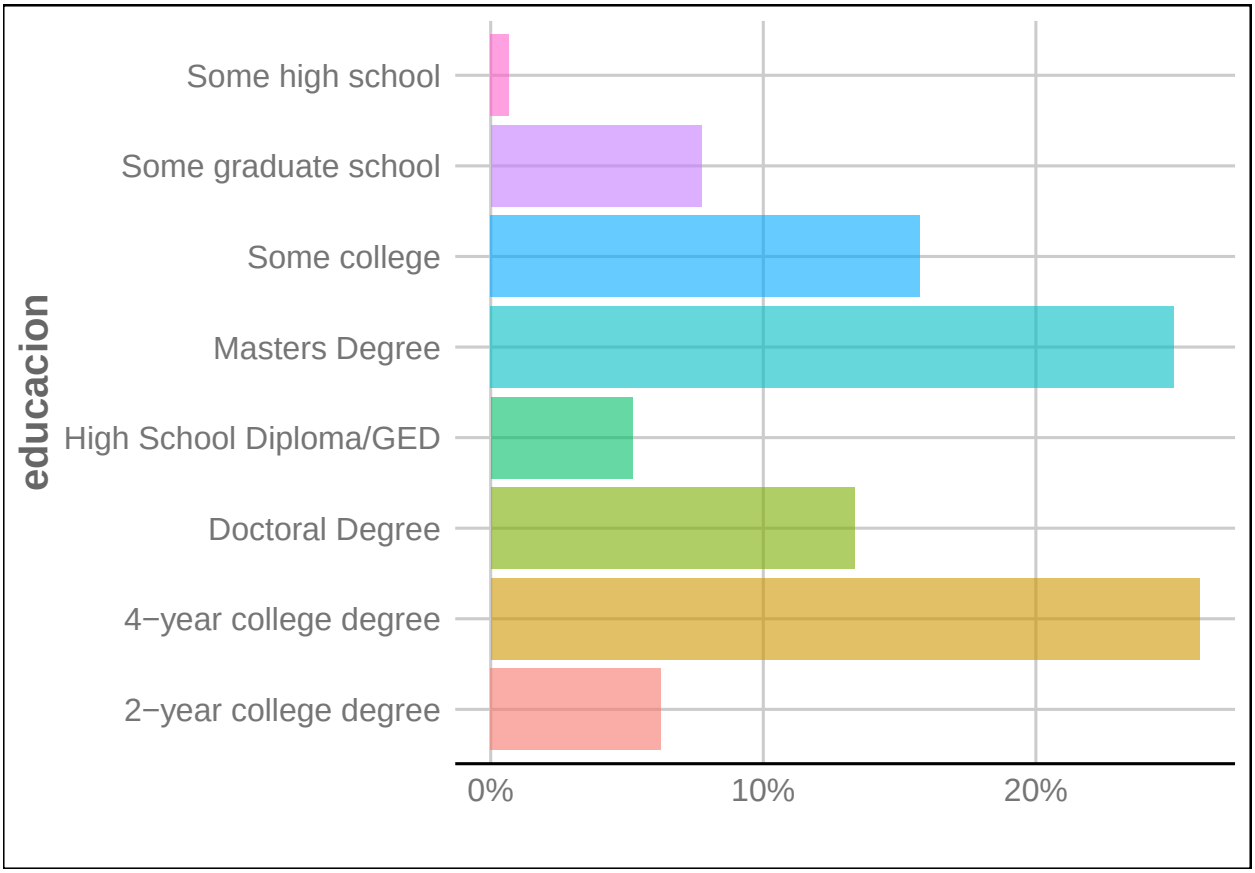
- En **cuidado** más del 80% de las personas se respondieron falso, casi el 5% si.

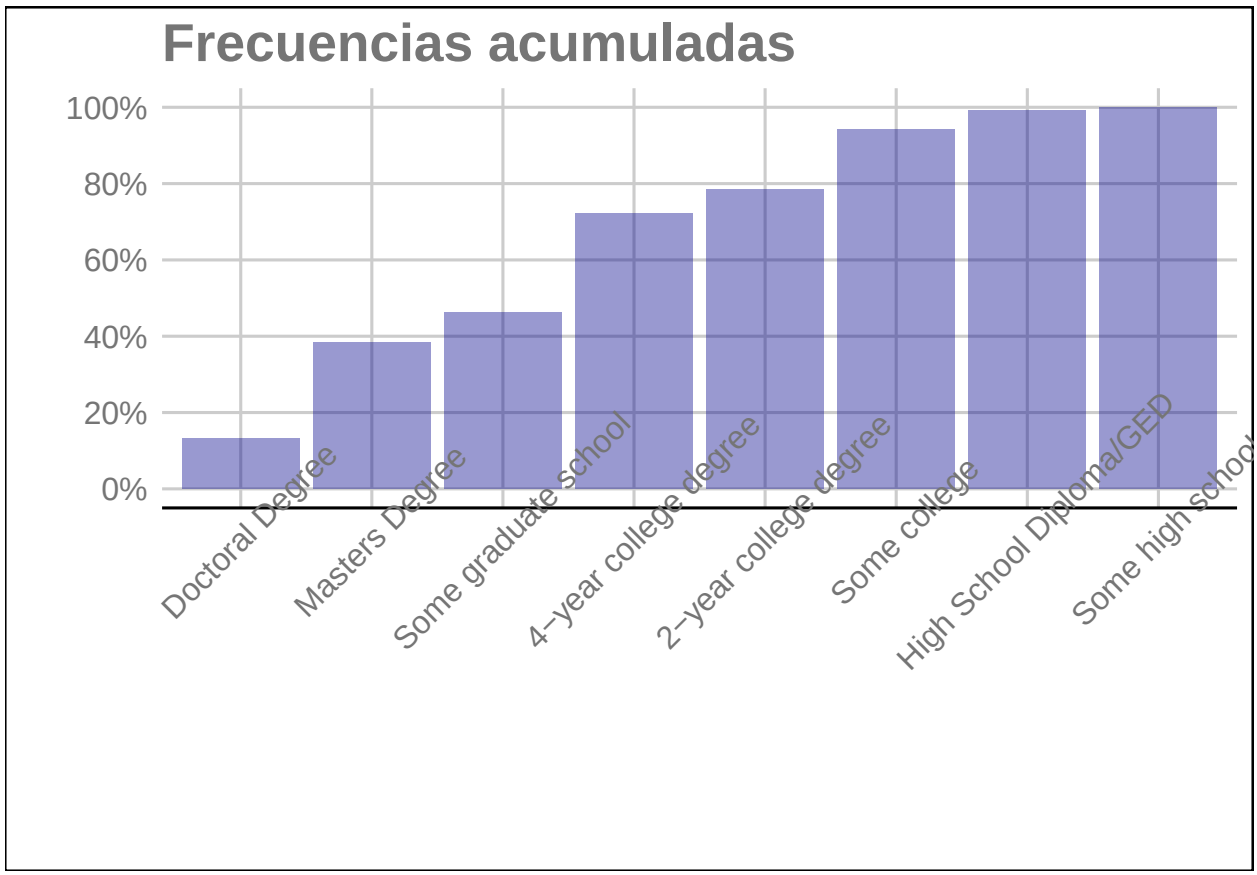
DBS	Freq
false	1850
true	100



- En **estimulacion_cerebral** más del 80% de los datos respondieron fallos y el resto verdadero.

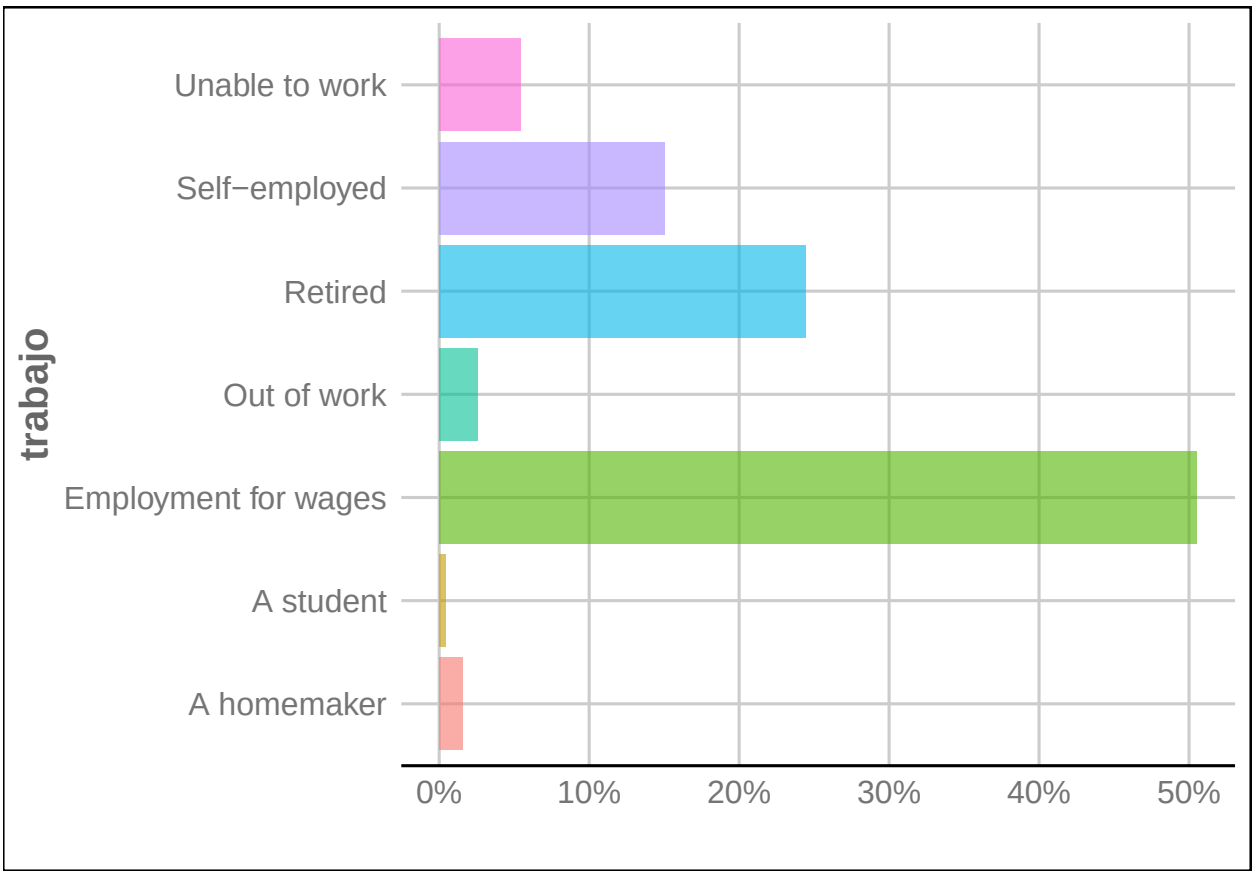
educacion	Freq
2-year college degree	138
4-year college degree	575
Doctoral Degree	295
High School Diploma/GED	115
Masters Degree	554
Some college	348
Some graduate school	171
Some high school	15


educacion
Some high school
Some graduate school
Some college
Masters Degree
High School Diploma/GED
Doctoral Degree
4−year college degree
2−year college degree
0%
10%
20%



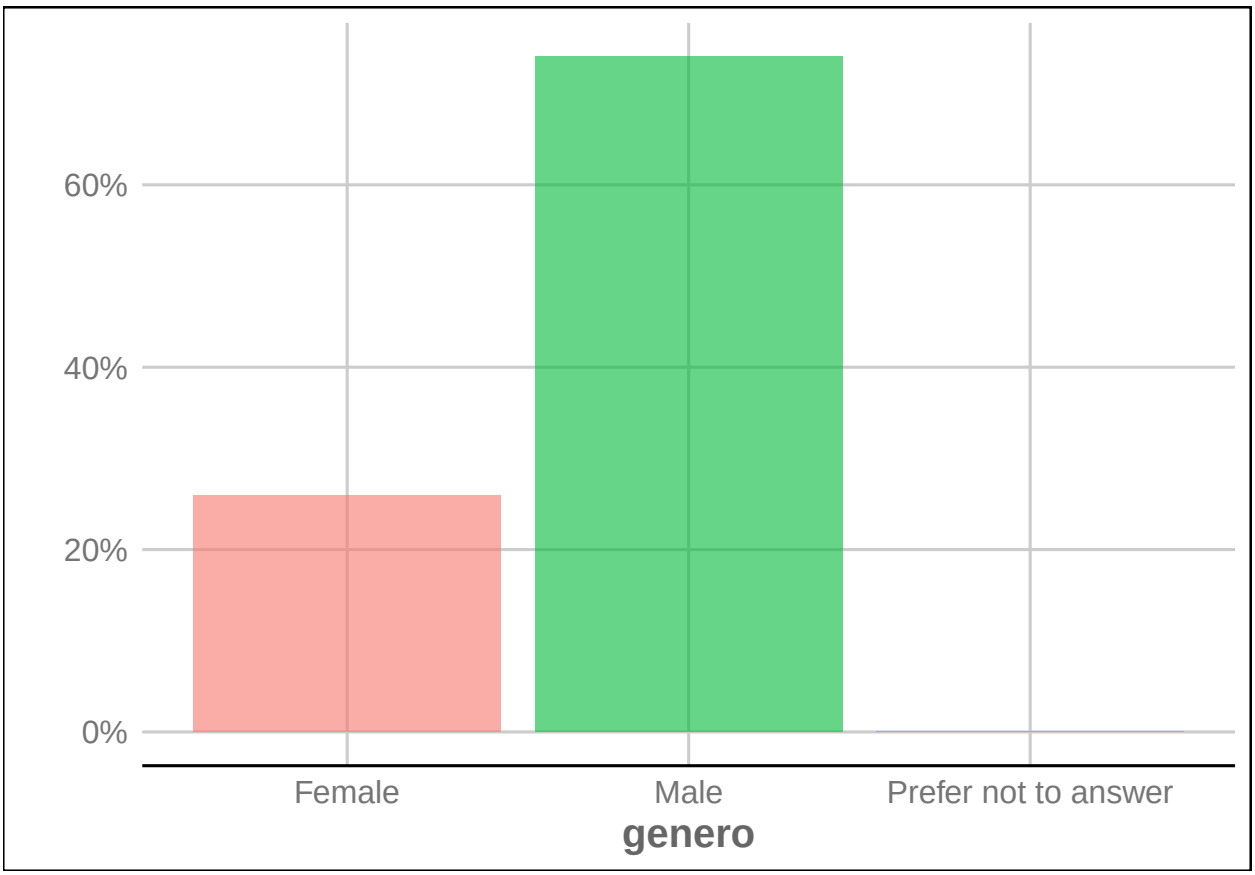
- En **educacion** se presenta la mayor cantidad de observaciones con "4-year college degree", seguido por "Masters Degree". En tercer lugar se tiene "Some college" y en cuarto "Doctoral Degree".
- Además, en el gráfico de barra se ordeno por el nivel de educación, para observar como son la distribución de los datos. Aquí encontramos que hasta "2-year college degree", es decir que tienen un título universitario, se acumula casi el 80% de los datos. Luego, si contamos "Some college" ya obtenemos más de los 90% de los datos.

trabajo	Freq
A homemaker	34
A student	10
Employment for wages	1117
Out of work	57
Retired	540
Self-employed	332
Unable to work	120



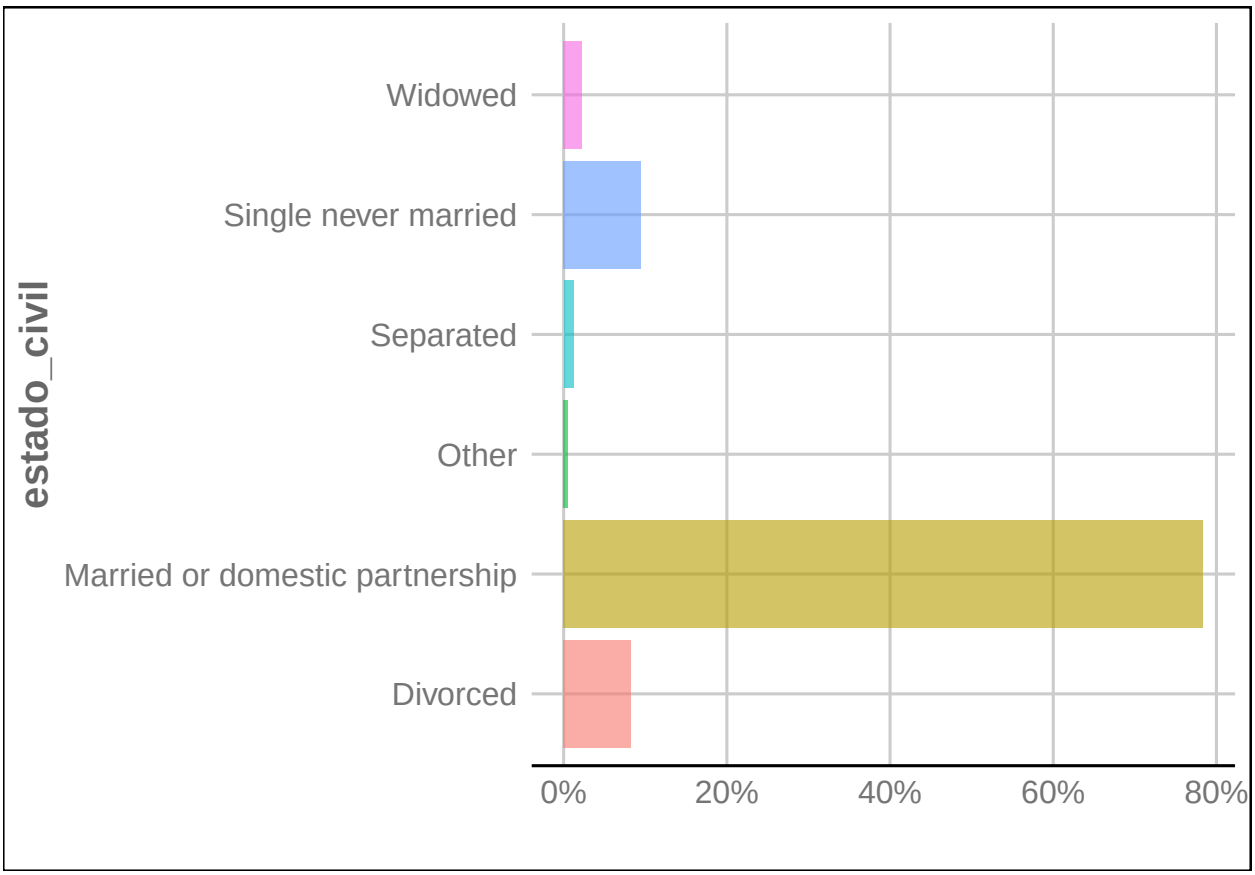
- En **trabajo** la mayor cantidad de los datos corresponden por “Empoyment for wages”, seguido por “Retired” y “Self-employed”.

genero	Freq
Female	574
Male	1641
Prefer not to answer	1



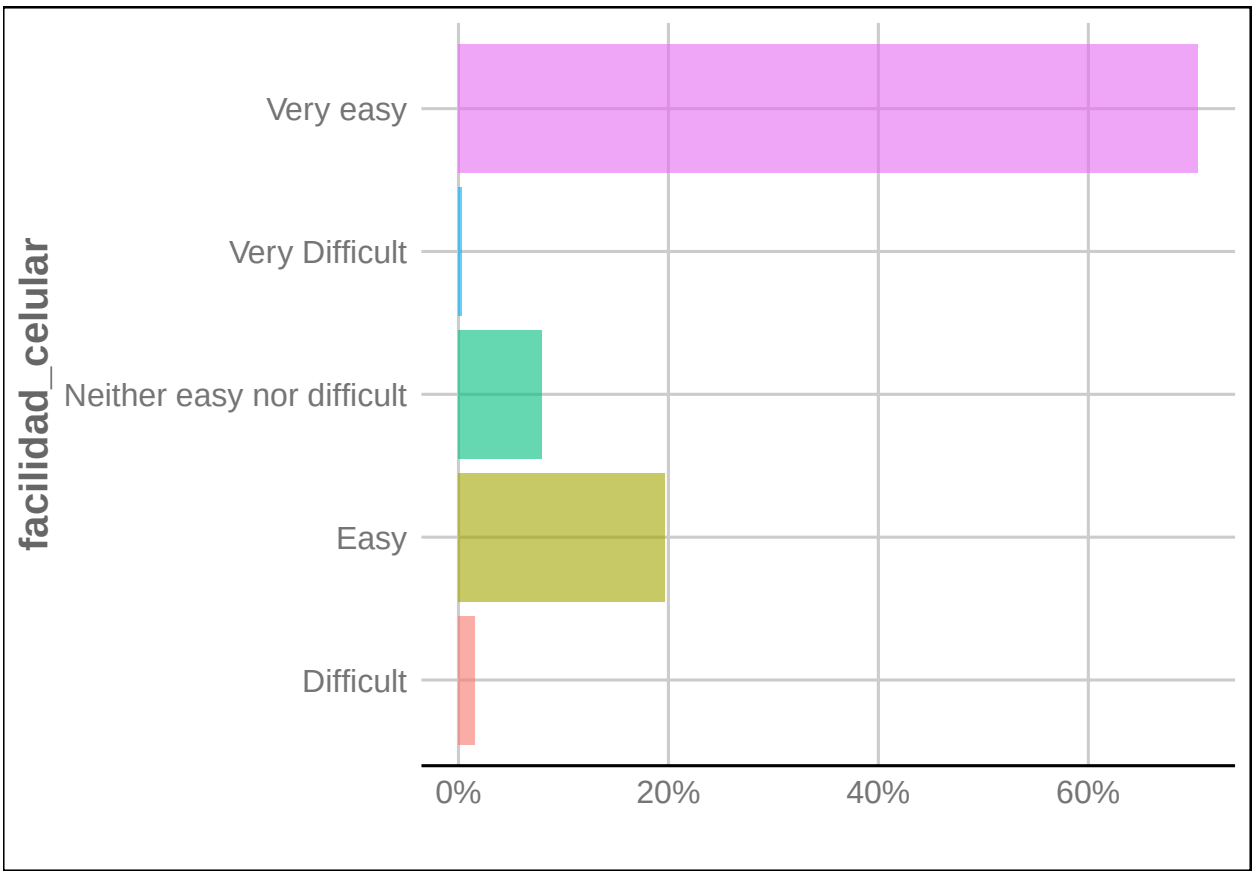
- En **genero** más del 70% de los datos corresponden a hombres, y un poco menos del 30% a mujeres.

estado_civil	Freq
Divorced	182
Married or domestic partnership	1735
Other	11
Separated	27
Single never married	209
Widowed	50



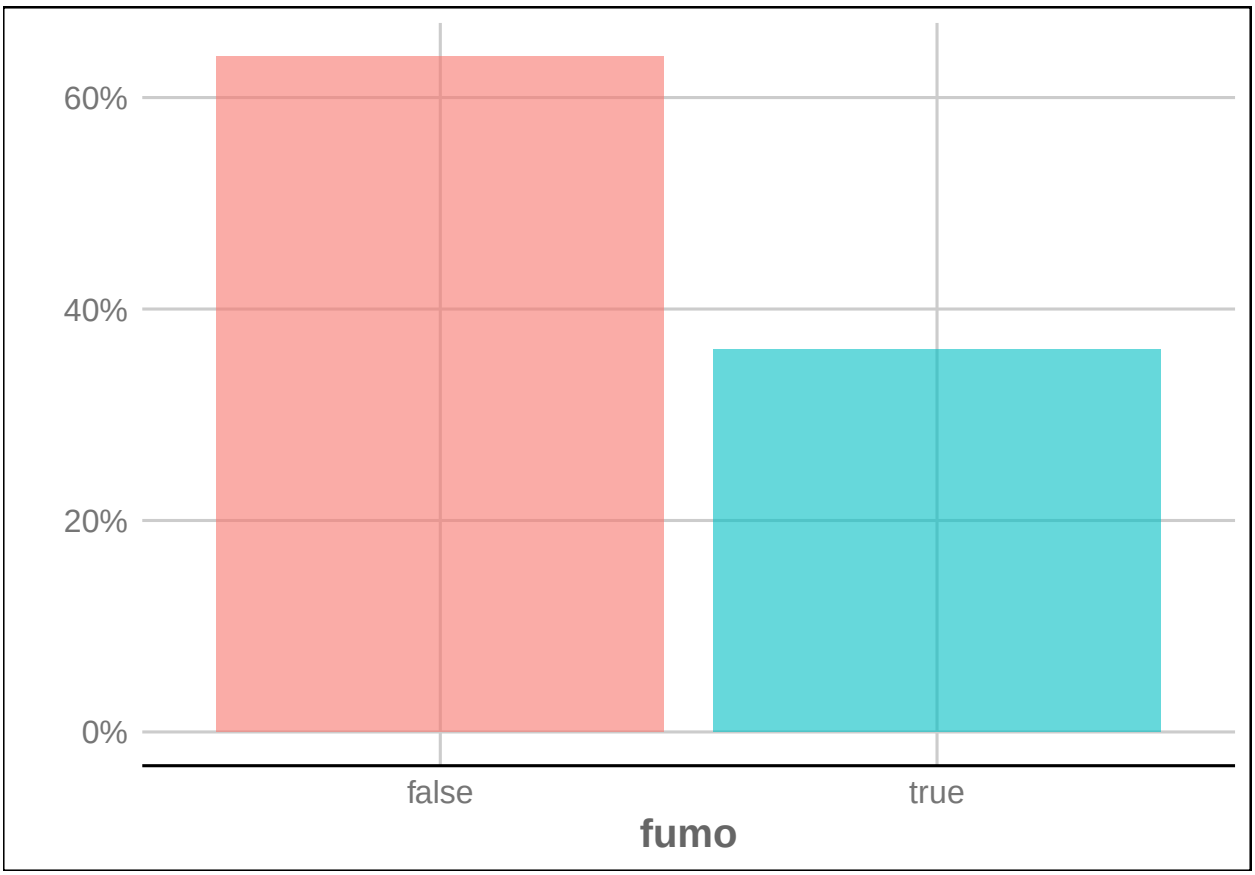
- En **estado_civil** casi el 80% de los datos se enceuntran en "Married or domestic partnership". Luego, siguen "Single never married" y "Divorced" con menos del 10% cada uno. El resto se distribuyo entre el resto de categorías.

facilidad_celular	Freq
Difficult	35
Easy	436
Neither easy nor difficult	176
Very Difficult	8
Very easy	1562



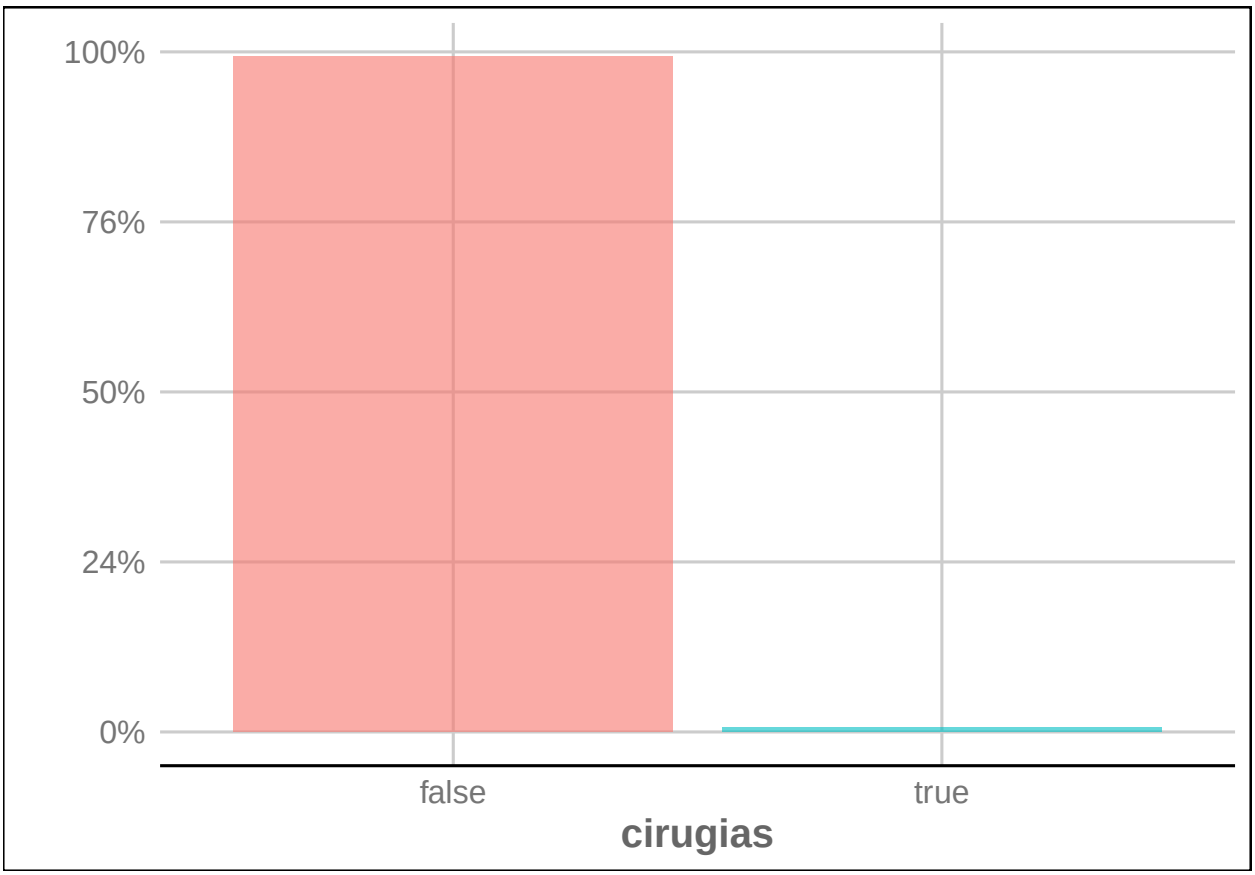
- En **facilidad_celular** el 70% de los datos coresponden a “Very easy”, un 20% a “Easy”, menos del 10% a “Neither easy nor dificult” y el resto a las demás categorías.

fumo	Freq
false	1375
true	778



- En **fumo** encontramos que más del 60% de los datos corresponden a falso y un poco más del 30% a verdadero.

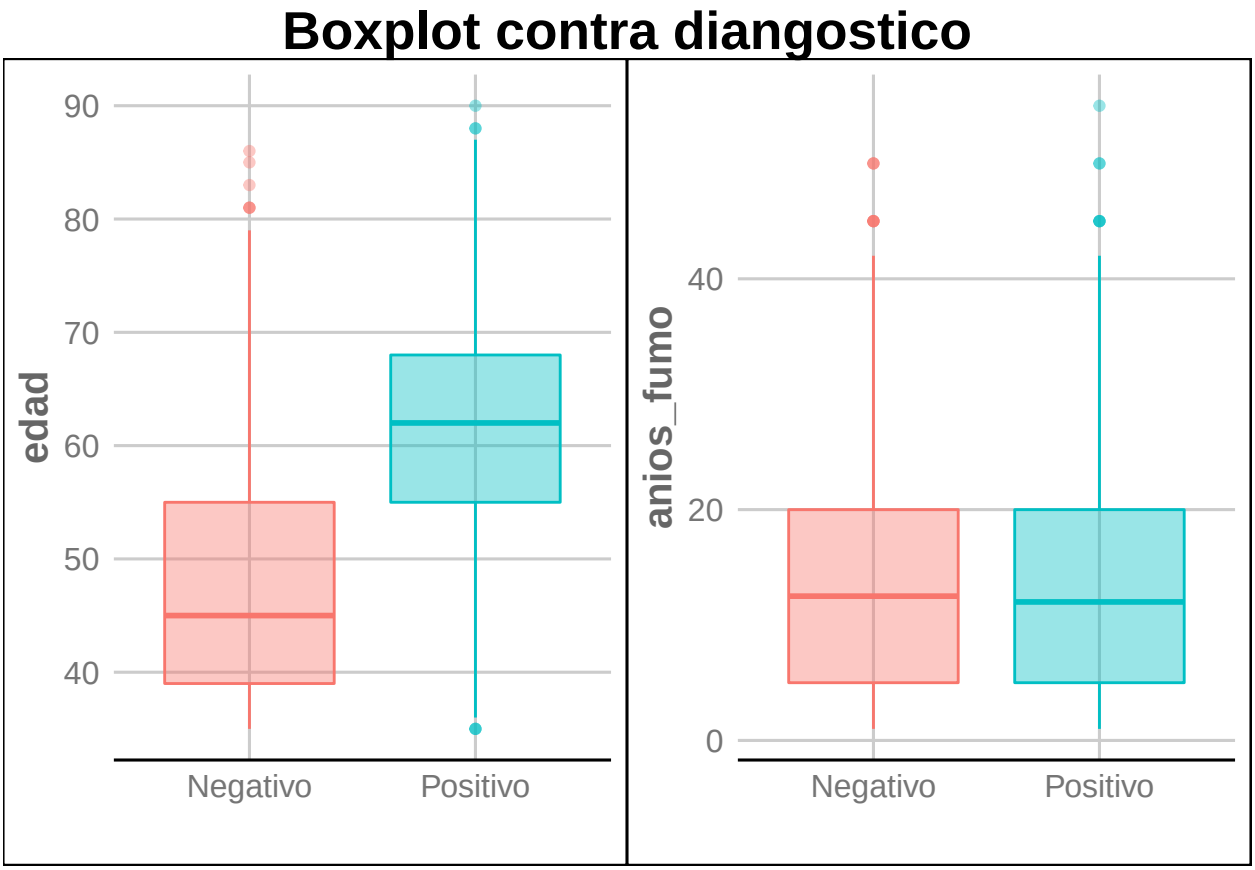
cirugias	Freq
false	1822
true	13



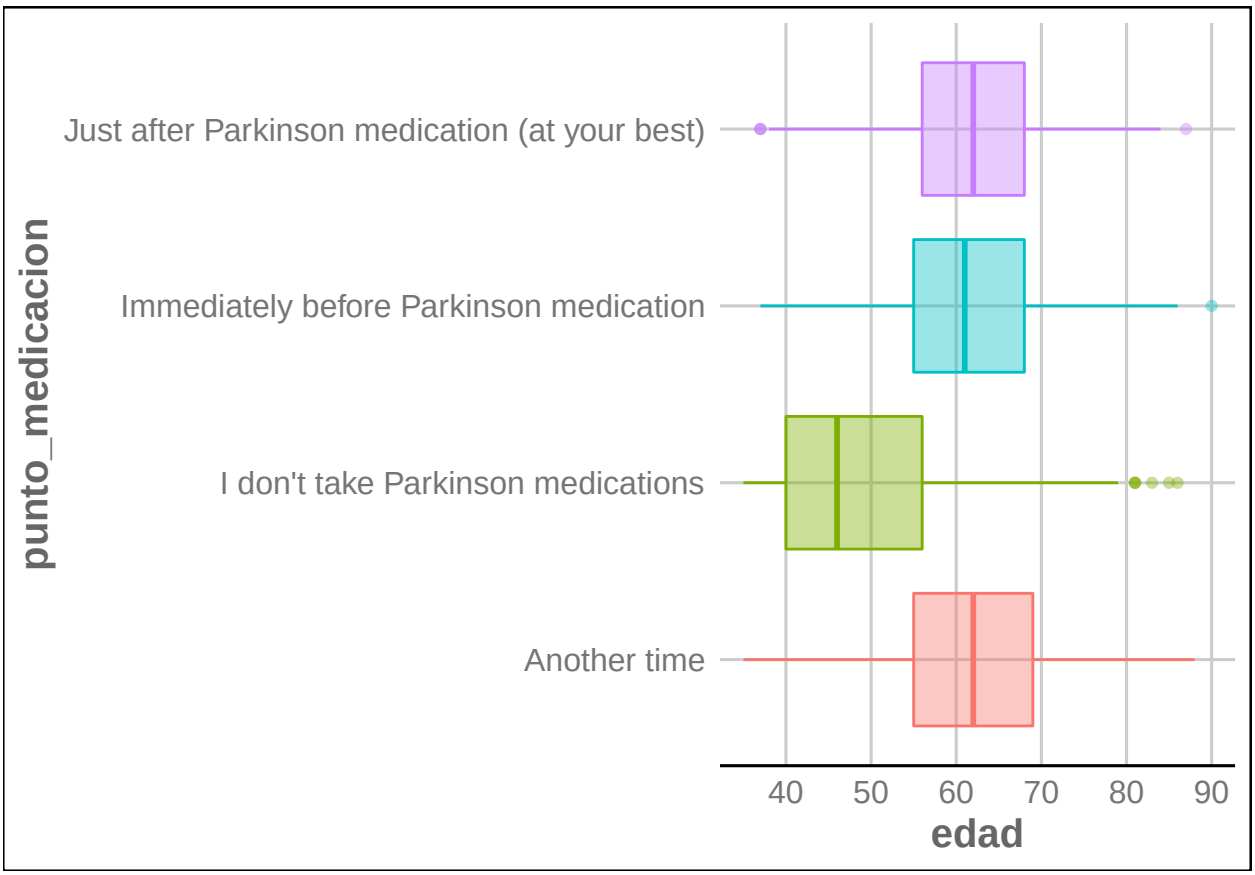
- En **cirugas** casi el 100% de los resultados corresponden a falos y un pequeño porcentaje a verdadero.

Análisis Multivariado

Cualitativa vs cuantitativa



- Se sugiere que la **edad** de las personas con diagnostico positivo es mayor a la edad con diagnostico negativo. Siendo la mediana de los primeros un poco superior a 60 años y 45 años para los segundos. Por el otro lado, **anios_fumo** no parece tener diferencia entre los paciente con diagnostico positivos y negativo.

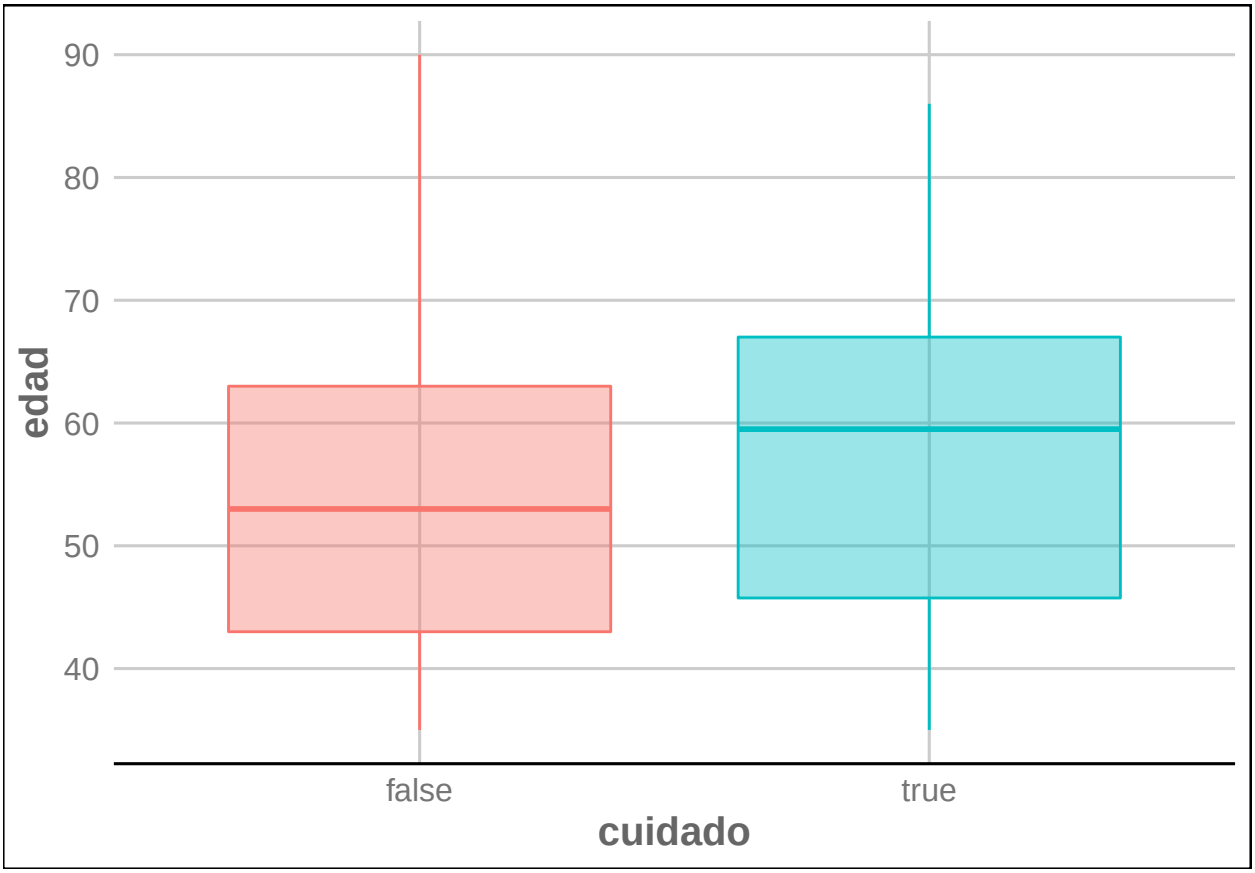


- Se observa como la **edad** parece ser menor en el caso de "I don't take Parkinsons medications" y los demás niveles son casi idénticos. Sin embargo, comparando con el gráfico previo de **edad** vs **diagnostico** se observa como las edades de las personas que tienen un diagnostico negativo son bastante similares a las edades de este grupo en particular. Entonces capaz que las personas de este grupo sean las que tienen un diagnostico negativo.

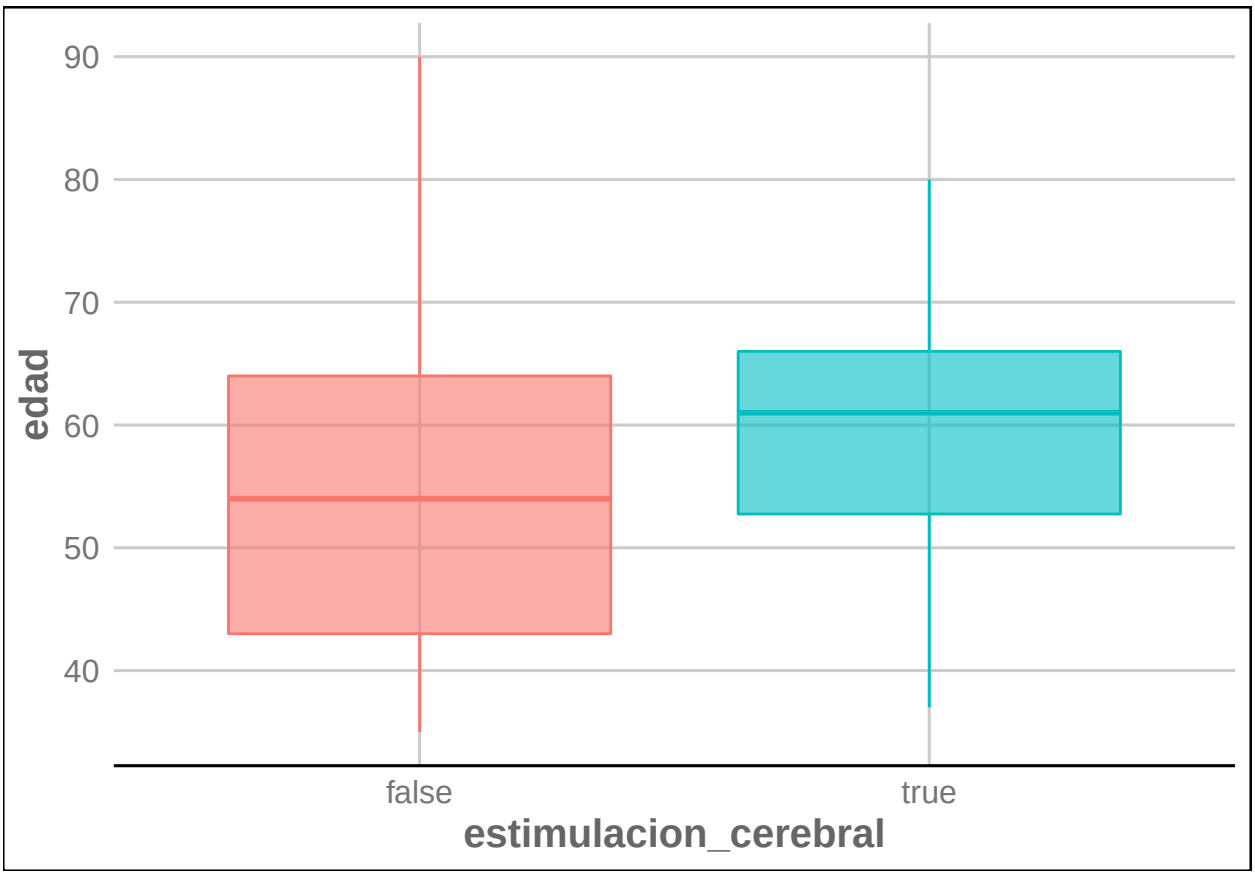
Realizamos la tabla de contingencia entre **diagnostico** contra **punto_medicacion** para verificar que las personas de "I don't take Parkinsons medications" son las que tienen un diagnostico negativo.

	0	1
Another time	0	394
I don't take Parkinson medications	1267	76
Immediately before Parkinson medication	0	209
Just after Parkinson medication (at your best)	0	255

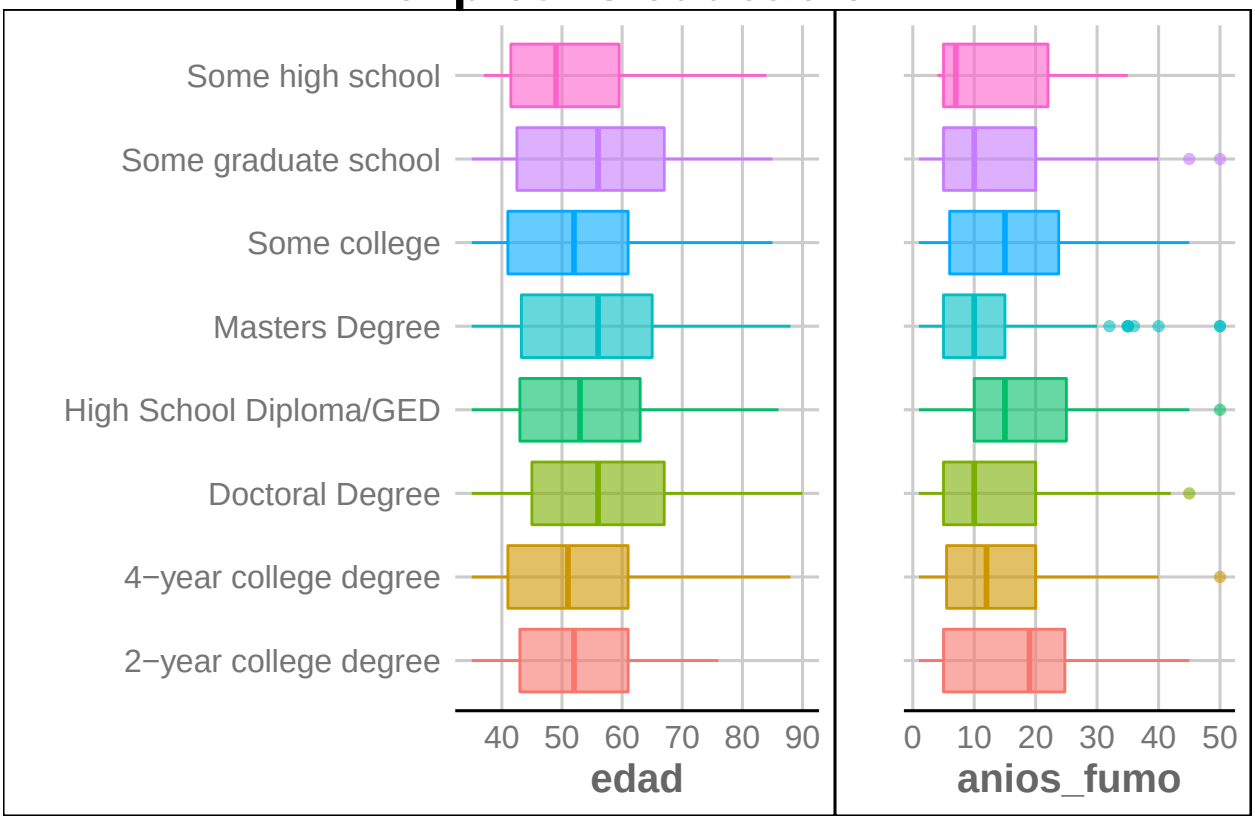
Aqui verificamos que la gran mayoría del segundo grupo tienen diagnóstico negativo (casi el 95%) y el resto positivo.



- No parece existir diferencia de edades entre los dos grupos.

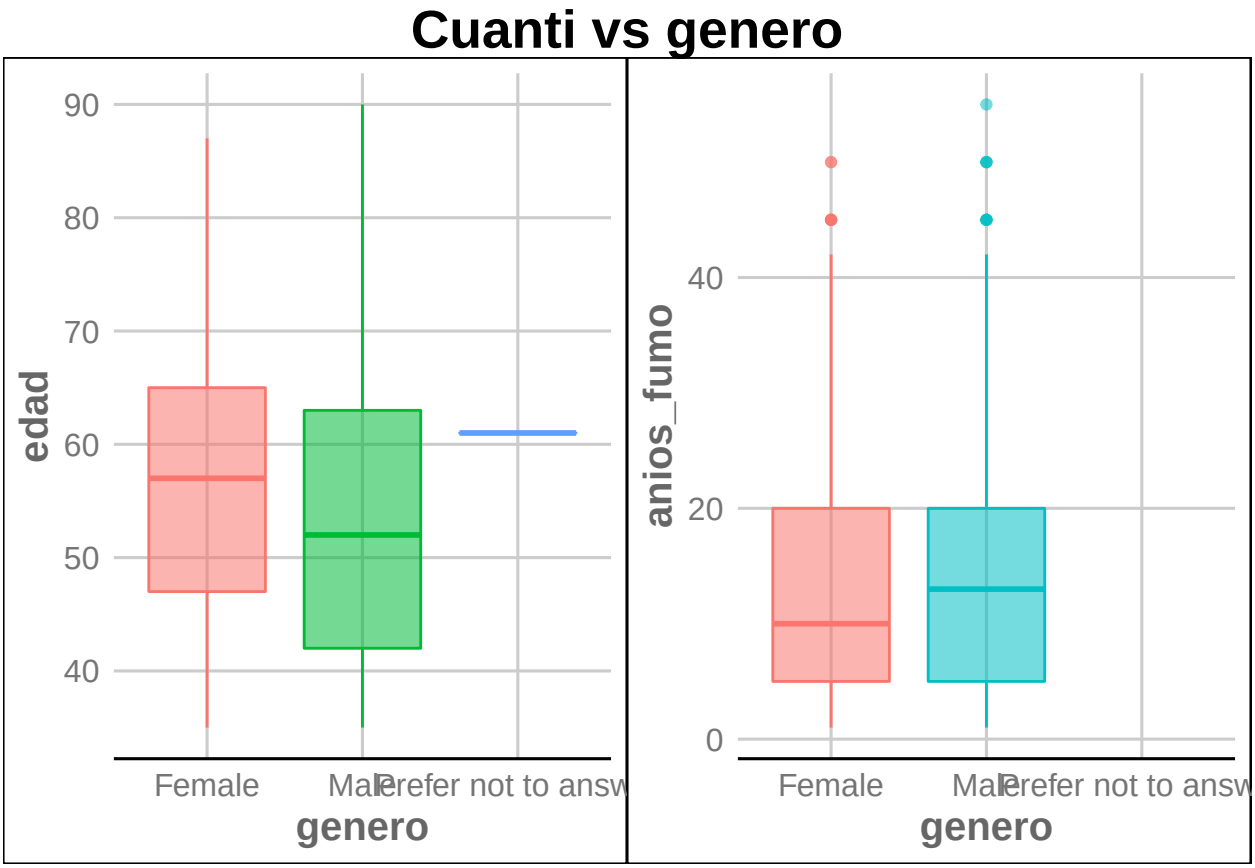


- La variabilidad de edades del grupo con el tratamiento deep brain parece ser menor que el otro grupo. Además, la edad mediana del grupo sin el tratamiento es menor al 75% del otro grupo (datos a partir del primer cuartil).

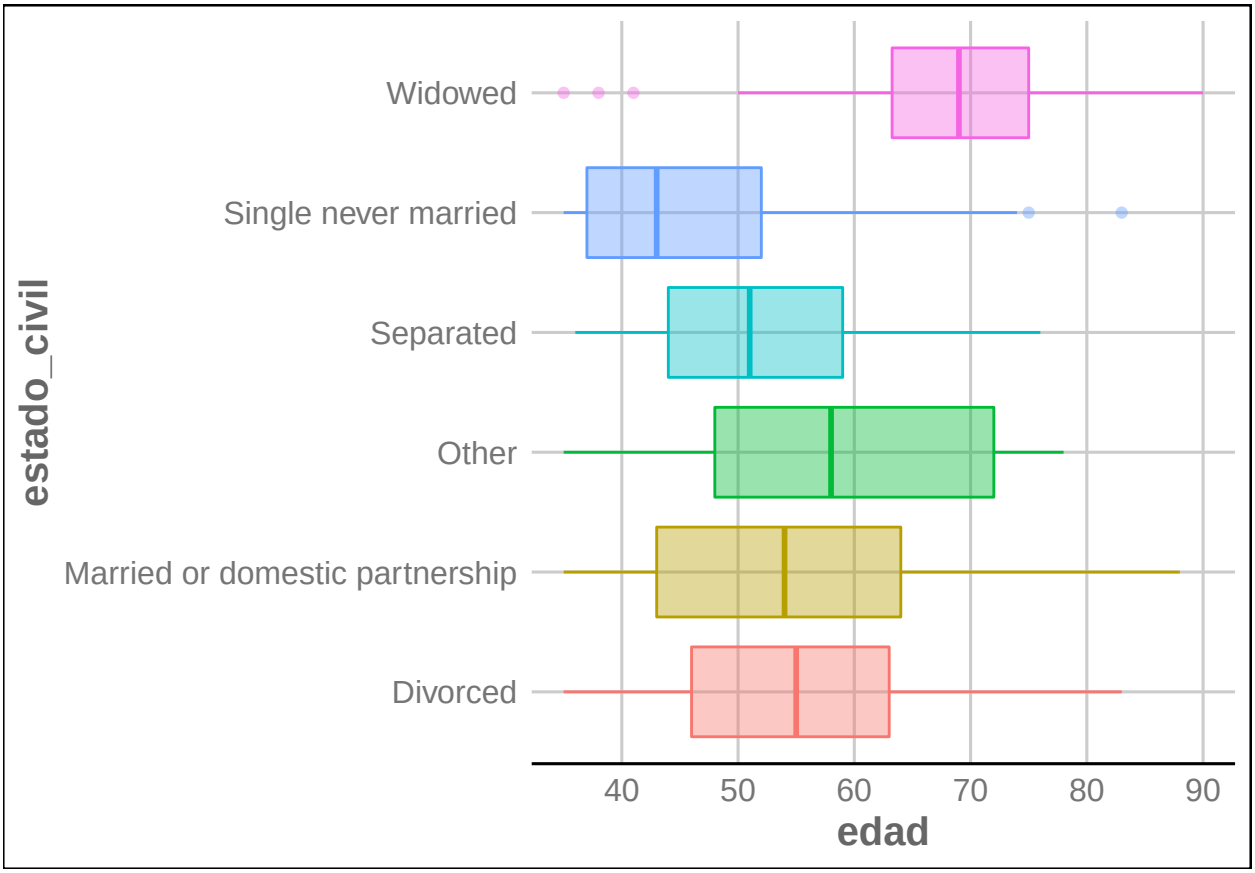


- En general, parece que todos los grupos presentan edades similares, aunque "Some graduate school" pareciera que la variabilidad es mayor.
- Para **anios_fumo** parece que es similar para todos los grupos, pero el grupo "Masters Degree" presenta menor variablidad. En contraste "2-year college degree" es el grupo con mayor variabilidad.

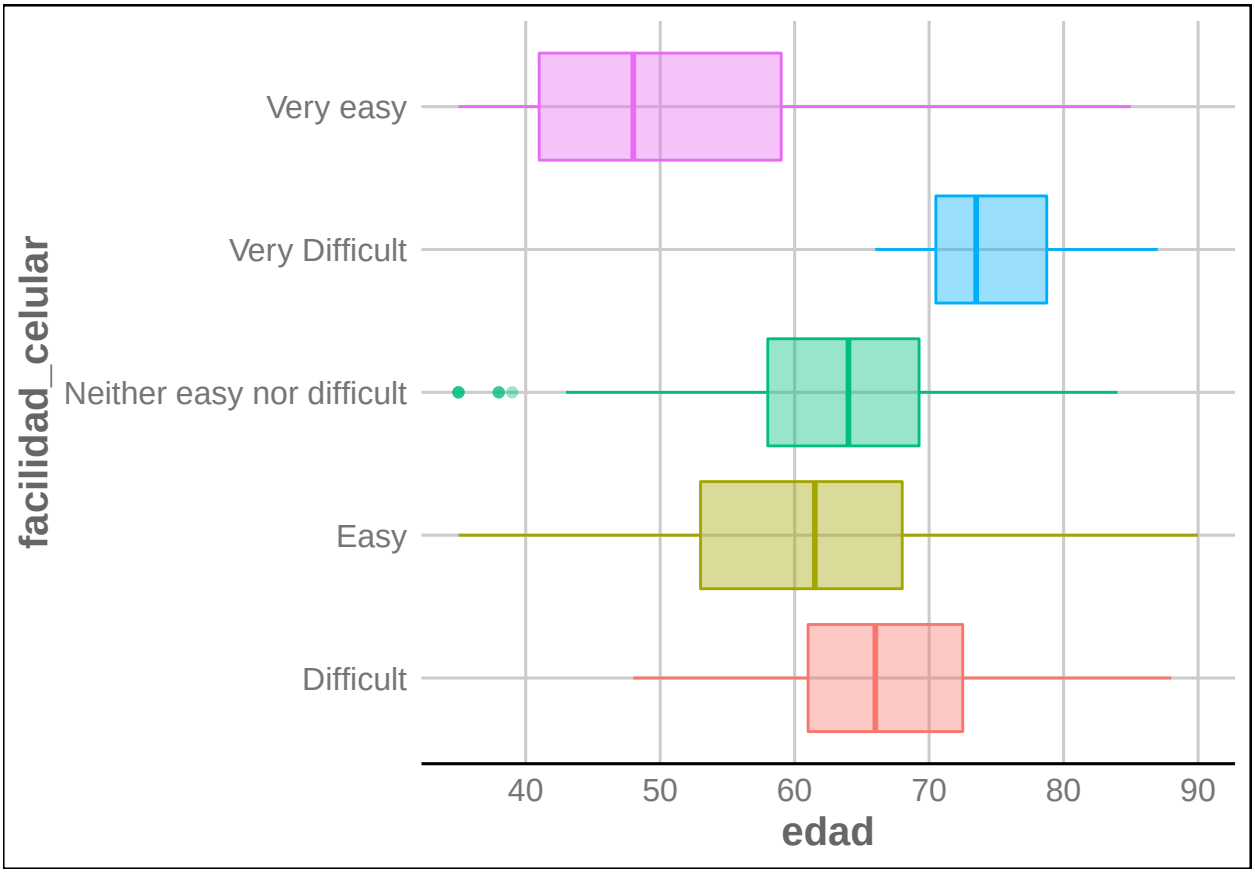
- El grupo de "A student" pareciera presentar edades menores al resto de grupos. A su vez "Retired" son los que tienen mayor edad, aunque encontramos varios valores extremos. El resto de grupos tienen edades intermedias, y en general presentan valores parecidos aunque la variabilidad cambia.
- En el segundo gráfico, todos los grupos presentan una cantidad de años que fumo similares. Aunque el grupo "A student" presenta edades menores al resto, **anios_fumo** es similar al de resto de grupos.



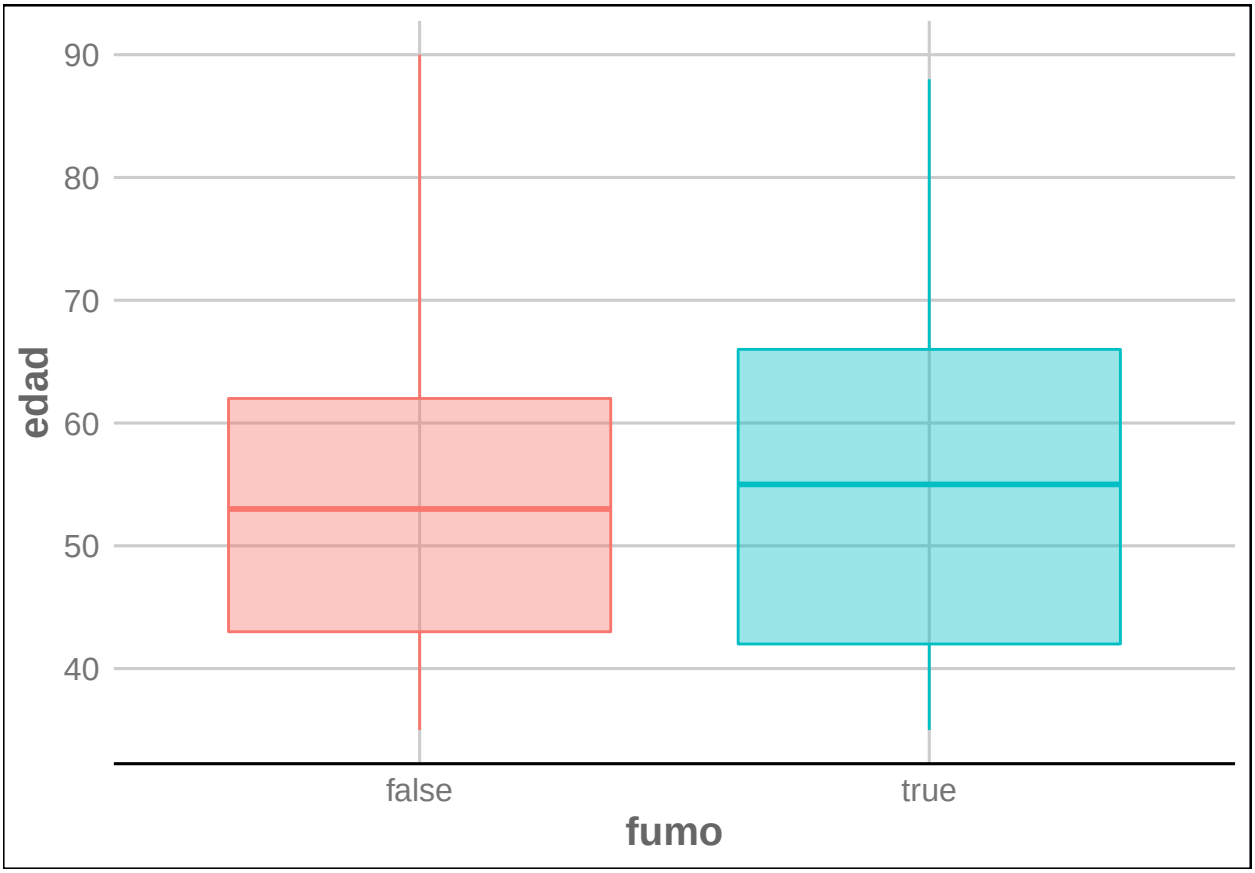
- No parece haber diferencia entre hombres y mueres en cuanto a la **edad** y **anios_fumo**



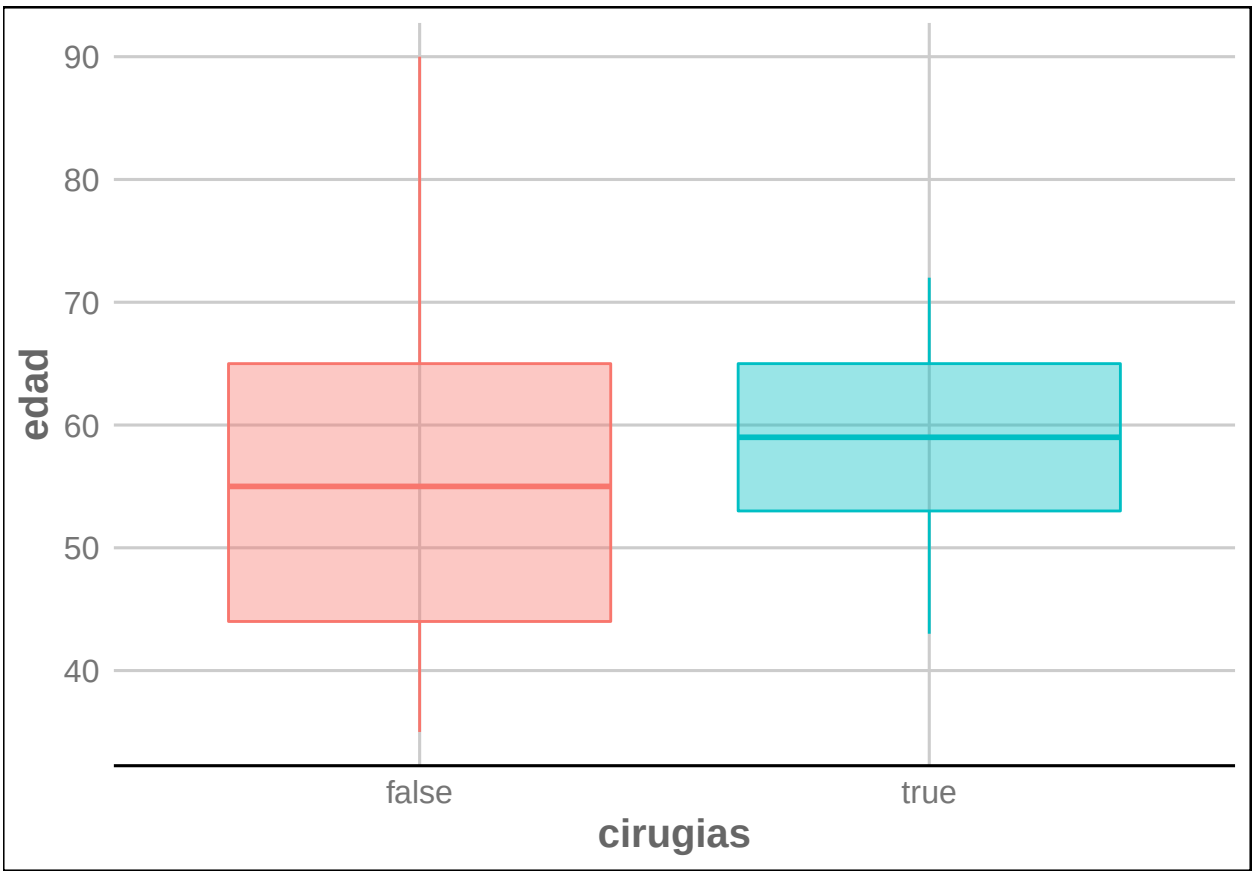
- El grupo “Single never married” presenta una edad mediana menor al primer cuartil del resto de los grupos. A su vez el grupo “widowed” presenta el primer cuartil superior al tercer cuartil del resto de grupos, excepto “Other”.



- El grupo “Very easy” presenta una edad mediana menor al primer cuartil del resto de los grupos. En cambio, el grupo “Very Difficult” presenta una edad mediana mayor al tercer cuartil del resto de grupos. Sin embargo, los tres grupos restantes presentan edades similares, teniendo a los grupos de “Easy” y “Difficult” aqui.



- No parece haber diferencia de edades entre ambos grupos.



- No parece haber diferencias de edades entre ambos grupos, pero el grupo que si tiene cirugías presenta menor variabilidad que el otro.

Cualitativa vs Cualitativa

	0	1	Sum
false	1206	869	2075
true	70	70	140
Sum	1276	939	2215

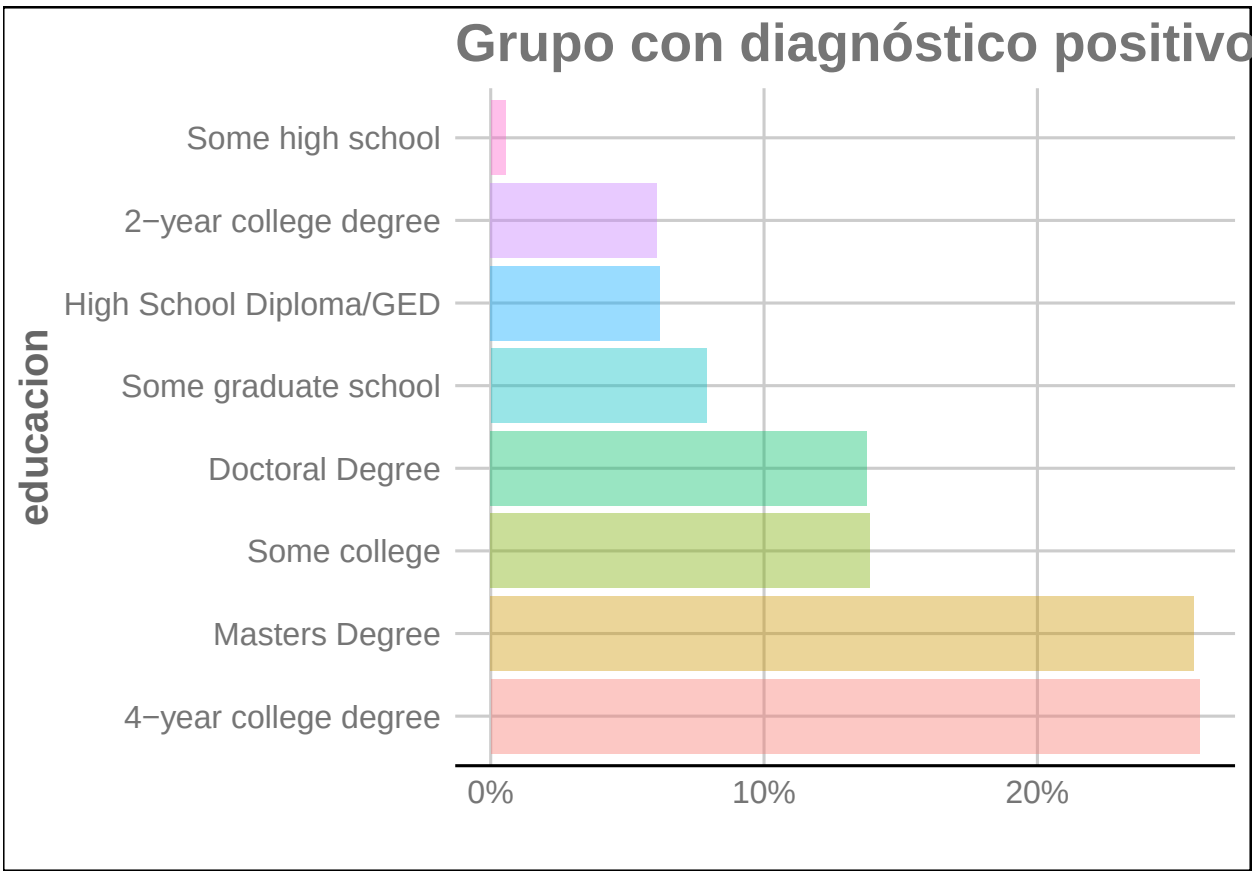
- Del grupo con diagnóstico negativo, el 94.45% no presenta cuidados y el resto si. Por le otro lado, para el grupo con diagnostico positivo presenta un 92.25% no presenta cuidado y el resto si.
- Del grupo que no tiene cuidado, el 58.12% presenta diangostico negativo, y el resto diagnóstico positivo. Por el otro lado, para el caso del grupo con cuidado, los datos se reparten exactamente a la mitad entre personas con diagnóstico positivo y negativo, es decir que hay un 50% para cada uno. Se observa como en estos casos la distribución de observaciones es más equitativa que cuando se miraban las columnas.

	0	1	Sum
false	1007	843	1850
true	6	94	100
Sum	1013	937	1950

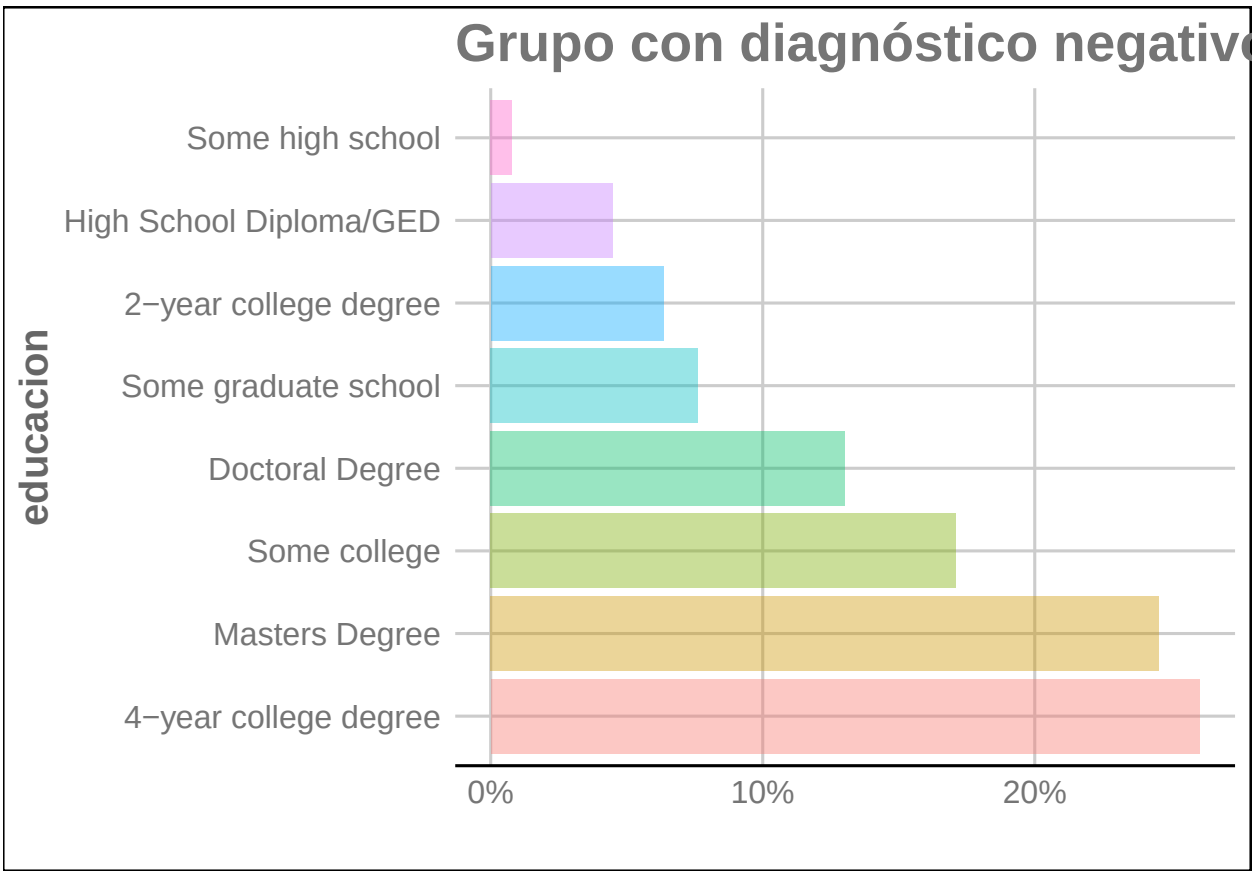
- Del grupo con diagnóstico negativo, el 99.4% no se realiza el tratamiento DBS y el resto si.
- Del grupo con diagnóstico positivo, el 89.96% no realiza el tratamiento de DBS y el resto sí.
- Del grupo que no realiza el tratamiento DBS, el 54.43% presenta diágnostico negativo y el resto positivo. La distribución de las osbservaciones es más equitativa en este caso.
- Del grupo que si realiza el tratamiento DBS, el 94% presenta diágnostico postivo y el resto negativo. Esto puede tener causa en que este tratamiento es usado generalmente en pacientes de Parkinson.

	0	1	Sum
2-year college degree	81	57	138
4-year college degree	332	243	575
Doctoral Degree	166	129	295
High School Diploma/GED	57	58	115
Masters Degree	313	241	554
Some college	218	130	348
Some graduate school	97	74	171
Some high school	10	5	15
Sum	1274	937	2211

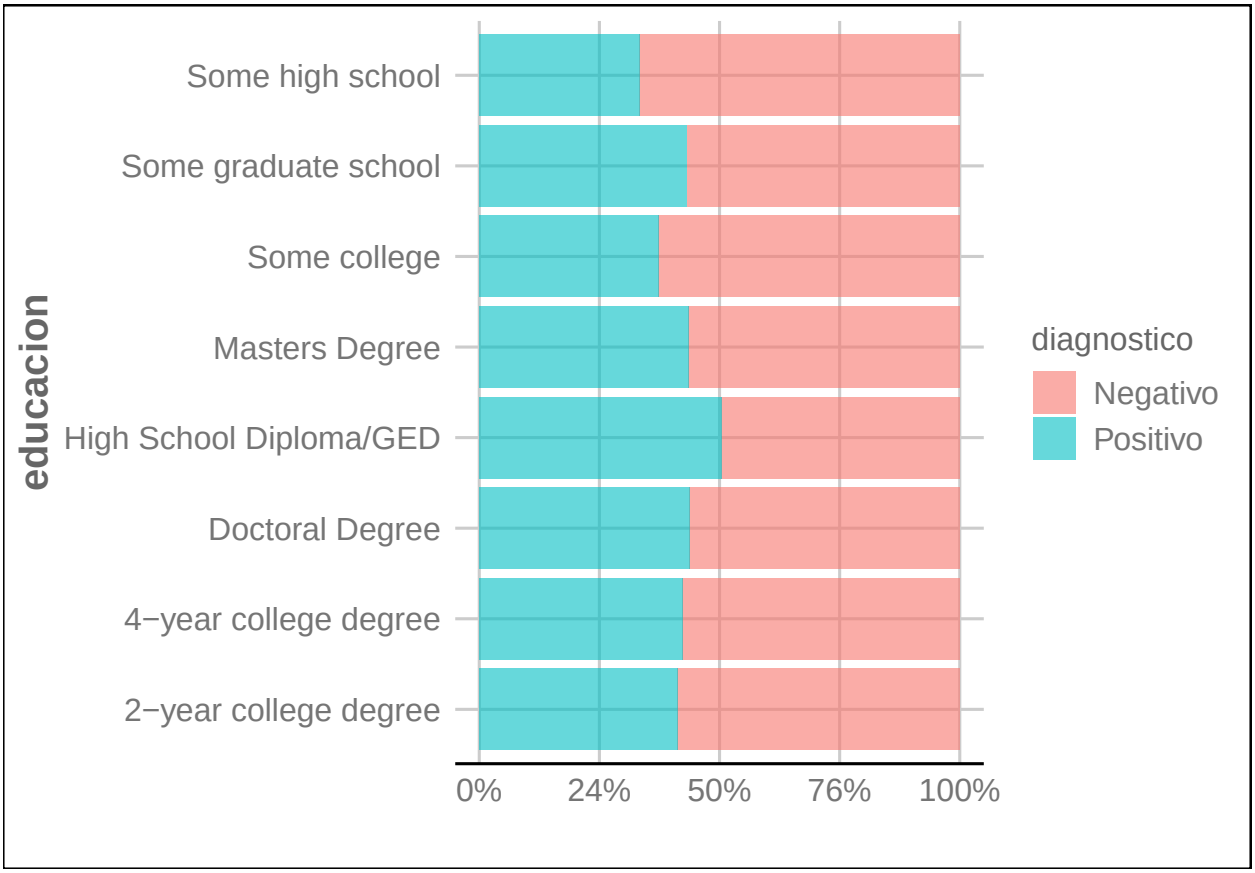
- En este caso, vamos a realizar dos diagramas de barras para mostrar como son la distribución de observaciones cuando se agrupa según el **diagnostico**



- Se observa como los grupos de "Masters Degree" y "4-year collegye degree" acumulan el 50% de las observaciones. Luego, entre el grupo de "Doctoral Degree" y "Some college" acumulan otro 20% más. El resto se reparte entre las categorias restantes.



- Se presentan resultados similares que en el caso de grupo con diagnóstico positivo.

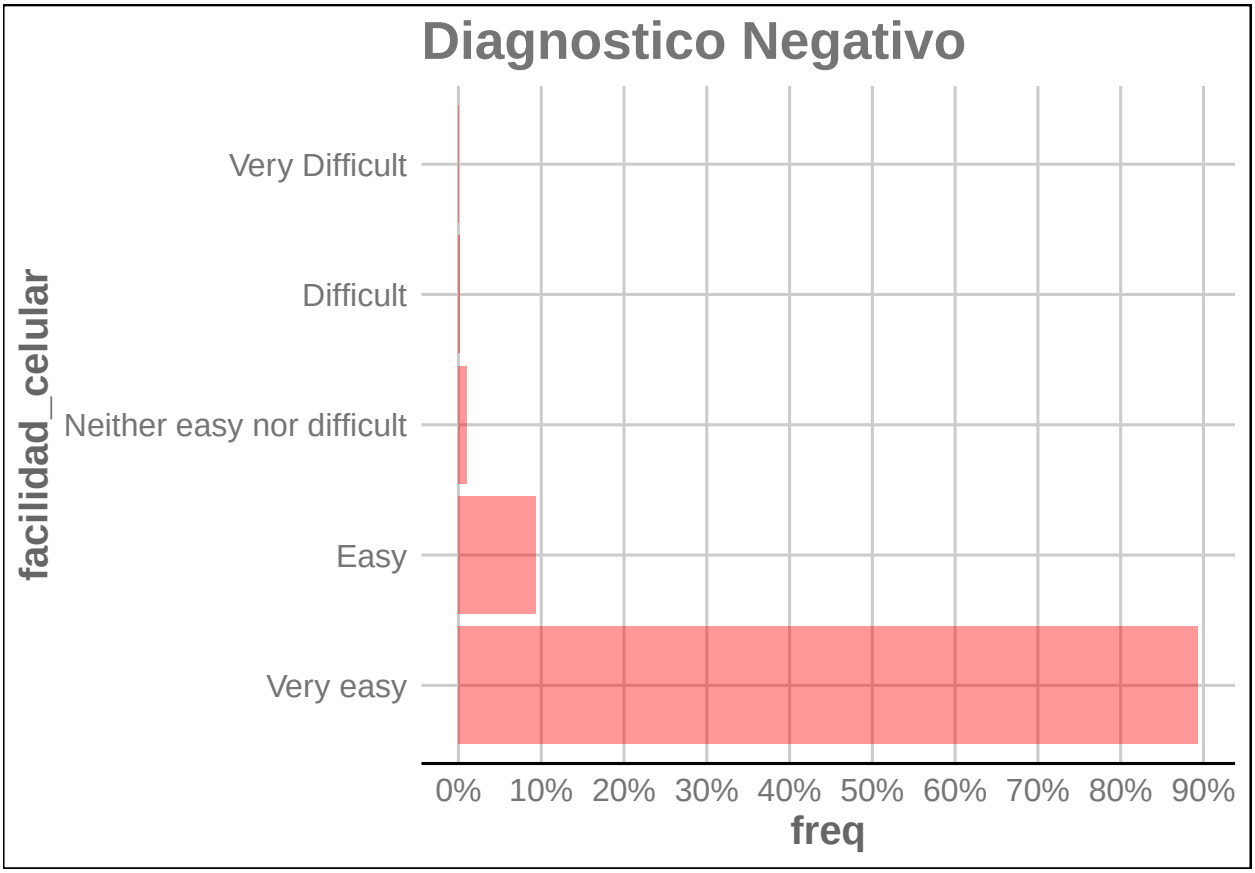


- Se observa como todos los niveles de educación presentan porcentajes parecidos de personas con porcentajes negativos/postivos. El caso de "High School Diploma" que tiene los datos repartidos equitativamente.

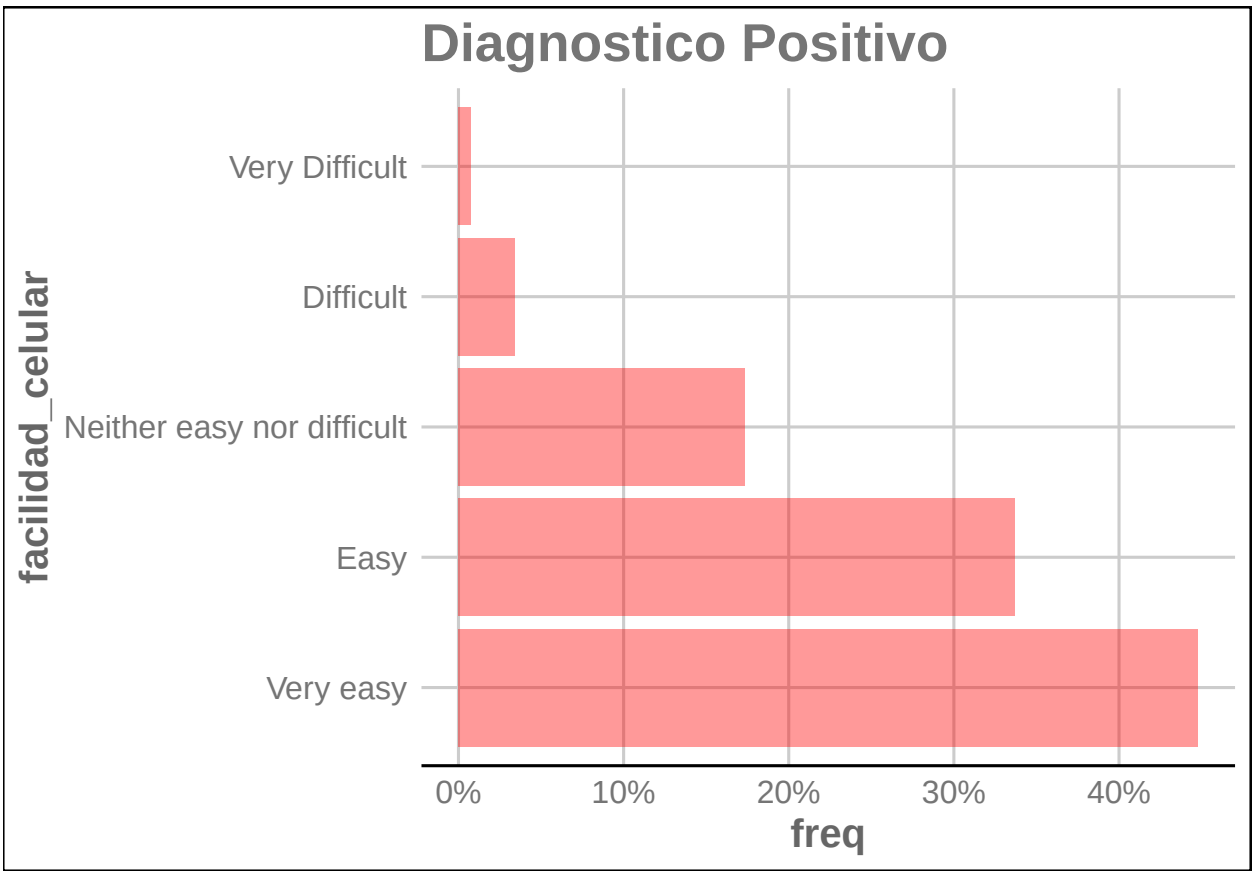
	0	1	Sum
Female	242	332	574
Male	1034	607	1641
Prefer not to answer	1	0	1
Sum	1277	939	2216

- Del grupo con diagnóstico negativo, el 19.73% son mujeres y el resto son hombres. Hay mayor presencias de hombres.
- Del grupo con diagnóstico postico, el 35.35% son mujeres y el resto hombres, al igual que el caso anterior hay mayor presencias de hombres.
- Del grupo de mujeres, el 42.16% tienen diagnóstico negativo y 57.89% postiivo. LOs datos se encuentran distribuidos de forma parecida en este caso.
- Del grupo de hombres, el 63.01 tiene diagnóstico negativo y 36.99% positivo. En este caso encontramos más hombres con diagnóstico negativo que positivos, mientras que en las mujeres era el caso inverso.

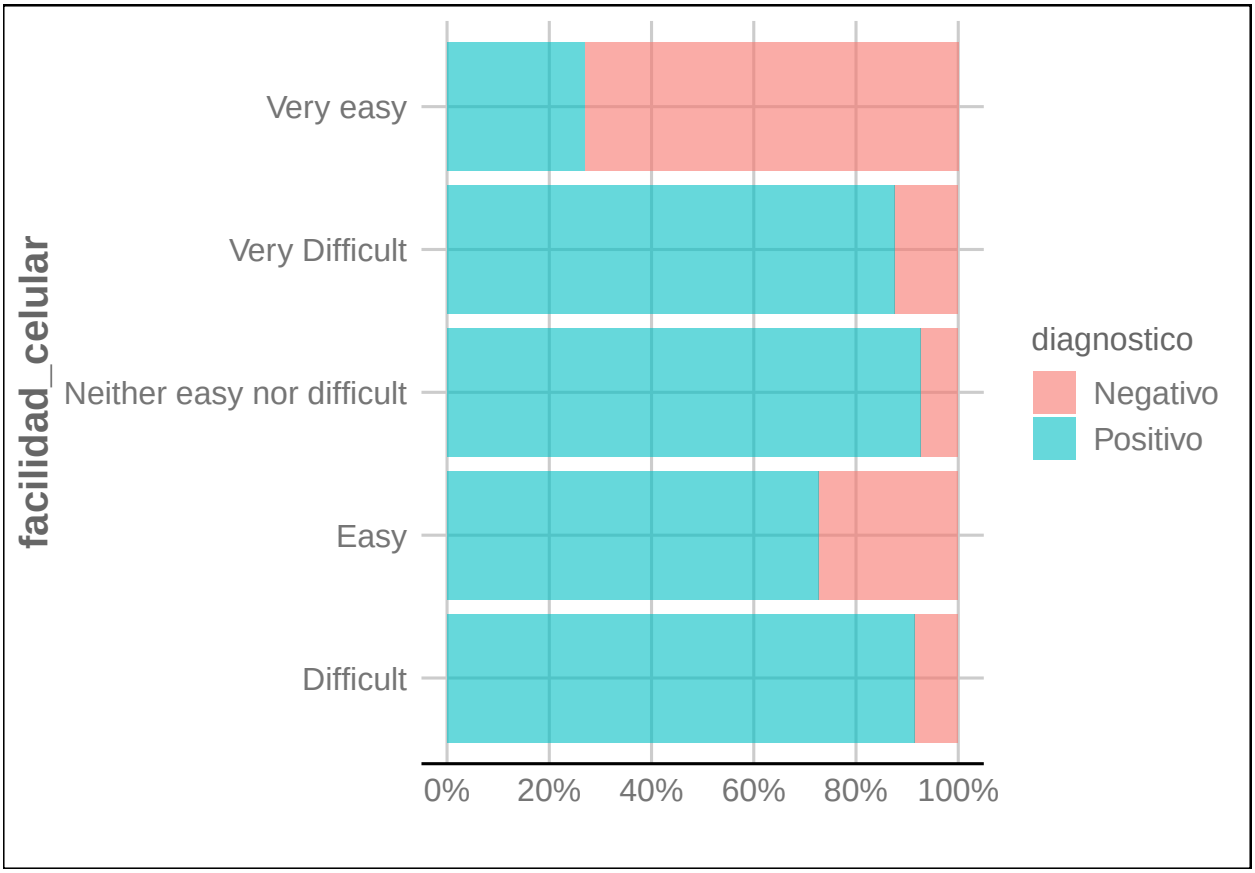
	0	1	Sum
Difficult	3	32	35
Easy	119	317	436
Neither easy nor difficult	13	163	176
Very Difficult	1	7	8
Very easy	1141	421	1562
Sum	1277	940	2217



- Se observa que casi el 90% de las personas con diagnóstico negativo se encuentran en “Very easy” y casi un 10% en “easy”.



- Si bien se presenta un orden de grupos idéntico al caso anterior, aquí los grupos “Neither easy nor difficult”, “Difficult” y “Very difficult” adquirien más relevancia en este grupo.



- Se observa como en todos los grupos, menos en "Very easy", se más del 60% de las personas con diagnóstico postivo. En "Very easy" ocurre lo contrario, la mayoría de las personas corresponden a diagnóstico negativo.

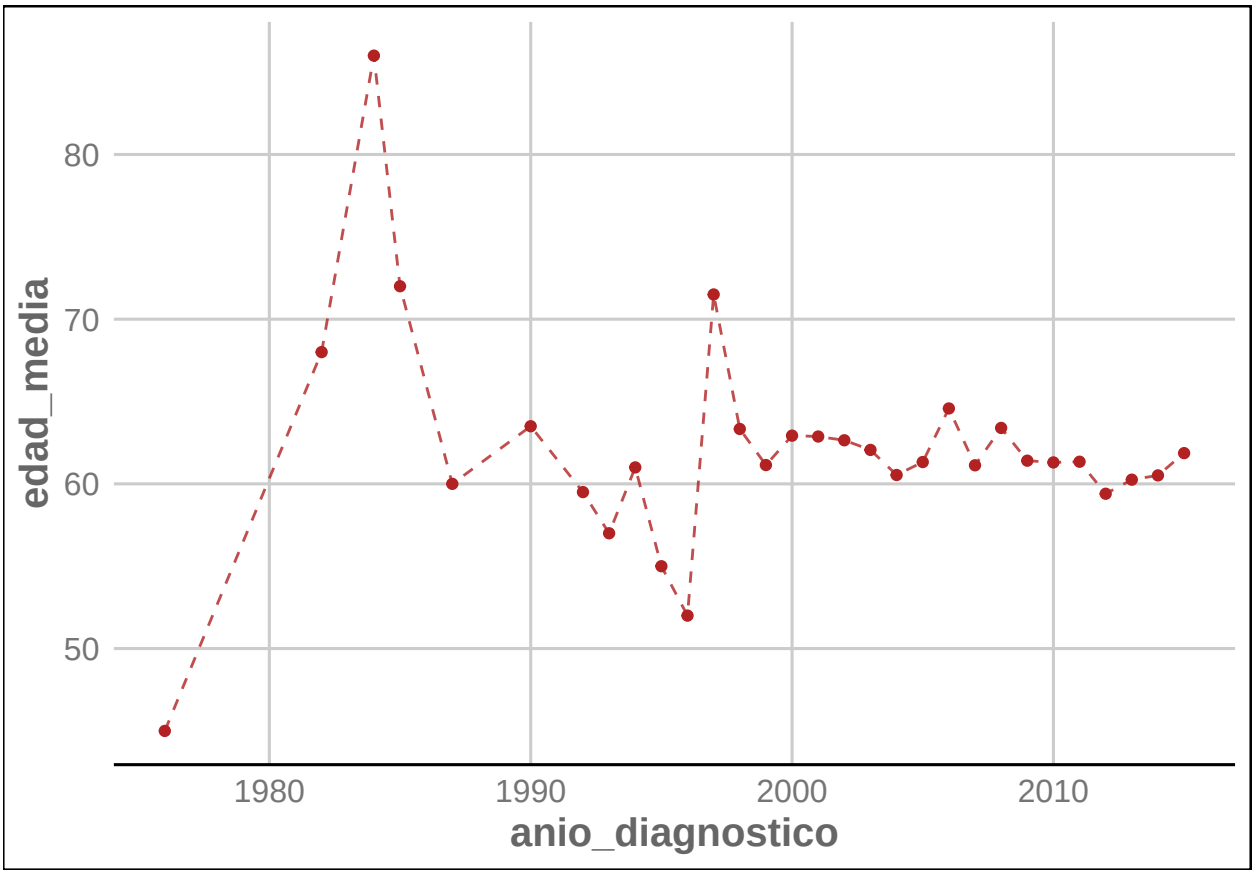
	0	1	Sum
false	759	616	1375
true	455	323	778
Sum	1214	939	2153

- Del grupo con diagnóstico negativo, el 62.52% no fumo y el 37.48% si.
- Del grupo con diagnóstico positivo, el 65.6% no fumo y el 34.4% si. Resultados similares al grupo anterior
- Del grupo que no fumo, el 55.2% tiene diagnóstico positivo y el 44.8% negativo. Encontramos que los datos estan distribuidos de forma equitativa.
- Del grupo que sí fumo, el 58.48% tiene diagnóstico positivo y el 41.52% negativo. Situación similar al caso anterior.

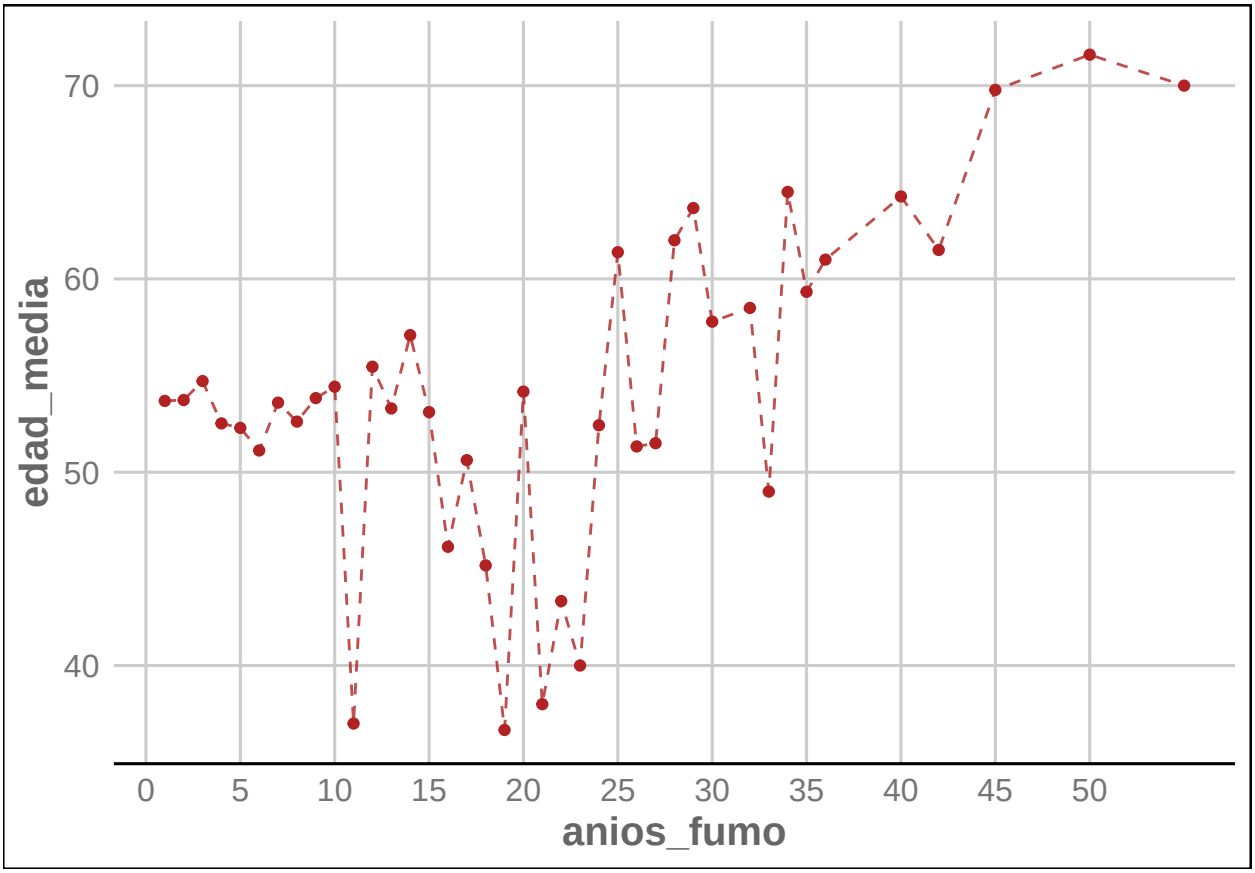
	0	1	Sum
false	894	928	1822
true	2	11	13
Sum	896	939	1835

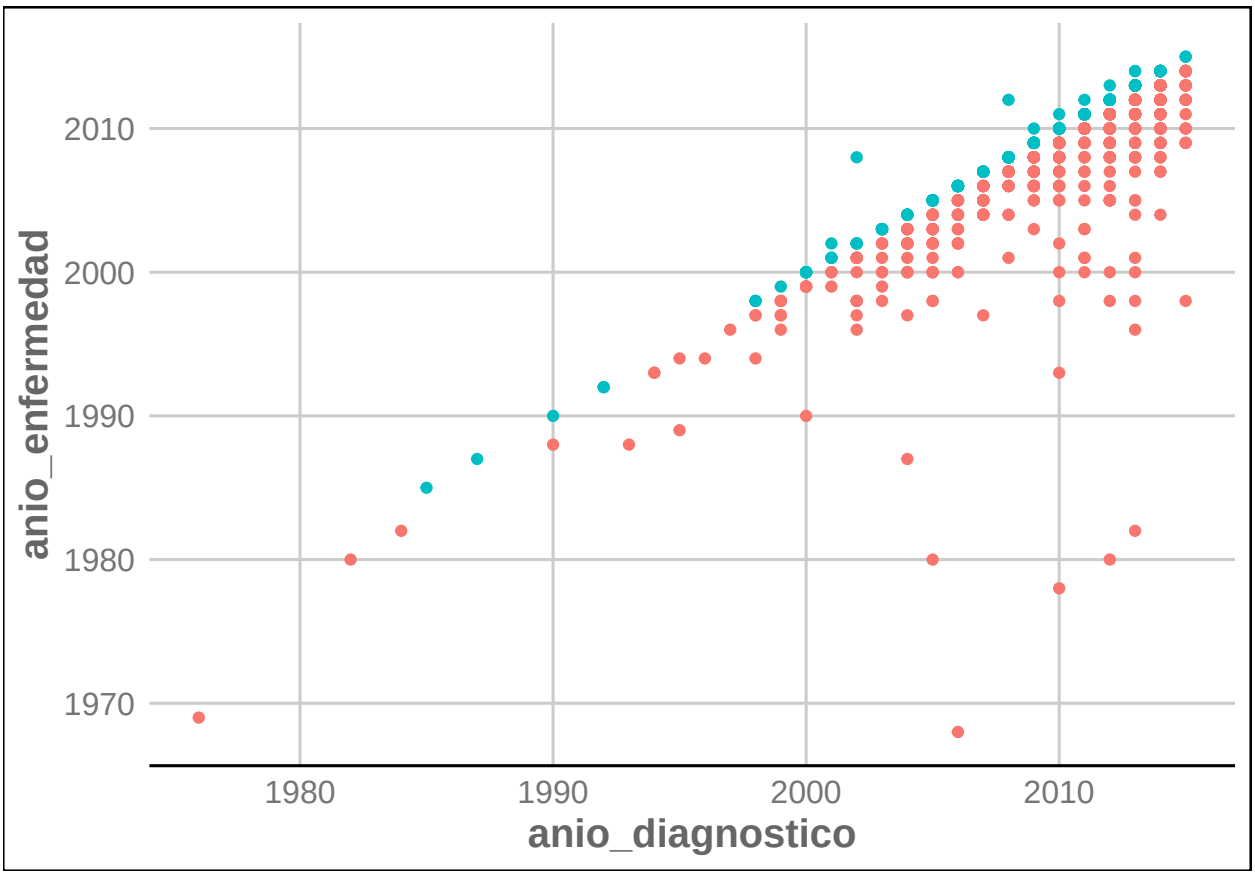
- Del grupo con diagnóstico negativo el 99.77% no presenta cirugías. En el grupo con diagnóstico positivo ocurre algo similar, hay un 98.82% de las personas sin cirugías.
- Del grupo sin cirugías, el 49.06% tienen diagnóstico negativo y el resto positivo. Se observa que los datos se encuentran repartidos equitativamente.
- Del grupo con cirugías, el 15.38% tienen diagnóstico negativo y el resto postivio. Aqui encontramos que predominan las personas con Parkinson.

Análisis cuanti vs cuanti



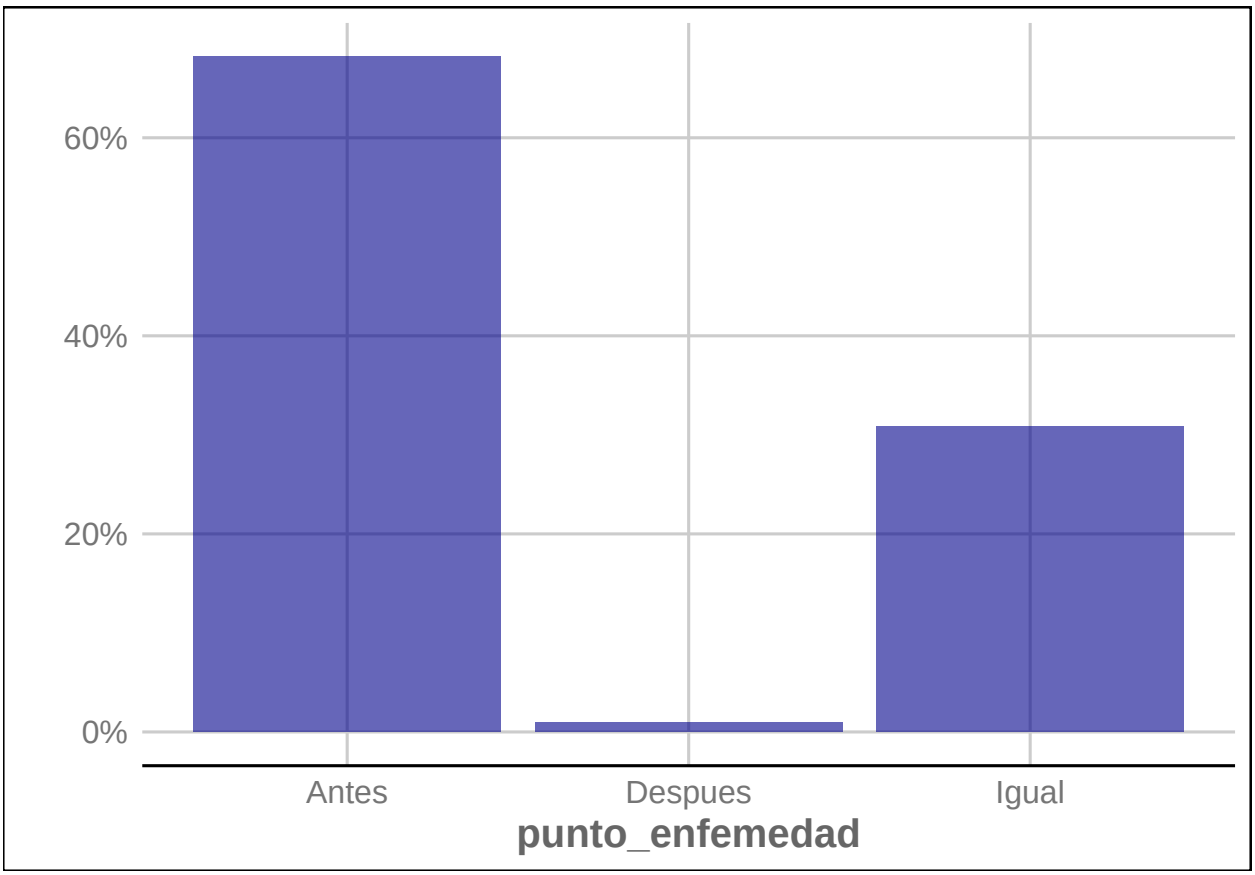
- Parece que al comienzo las edades medias varían considerablemente, desde casos con edades medias bajas o muy grandes. A partir del 200 parece que se estabiliza la edad media.


70
60
50
40
edad_media
0
5
10
15
20
25
30
35
40
45
50
anios_fumo



- Se observa que los puntos rojos pertenecen a los casos en que la persona recibió un diagnóstico luego de que se manisfestará la enferemdad.

Definimos una nueva variable lamada **punto_enfermedad** para boservar la cantidad de casos en que el diagnóstico ocurrió luego de que se manifestará la enfermedad.



- Más de la mitad de los casos recibieron el diagnóstico en años posetirores a que se manifesterá la enfermedad. En aproximadamente el 30% de los casos estos eventos ocurrieron en el mismo año. Y un pequeño procentaje primero recibio el diagnóstico.