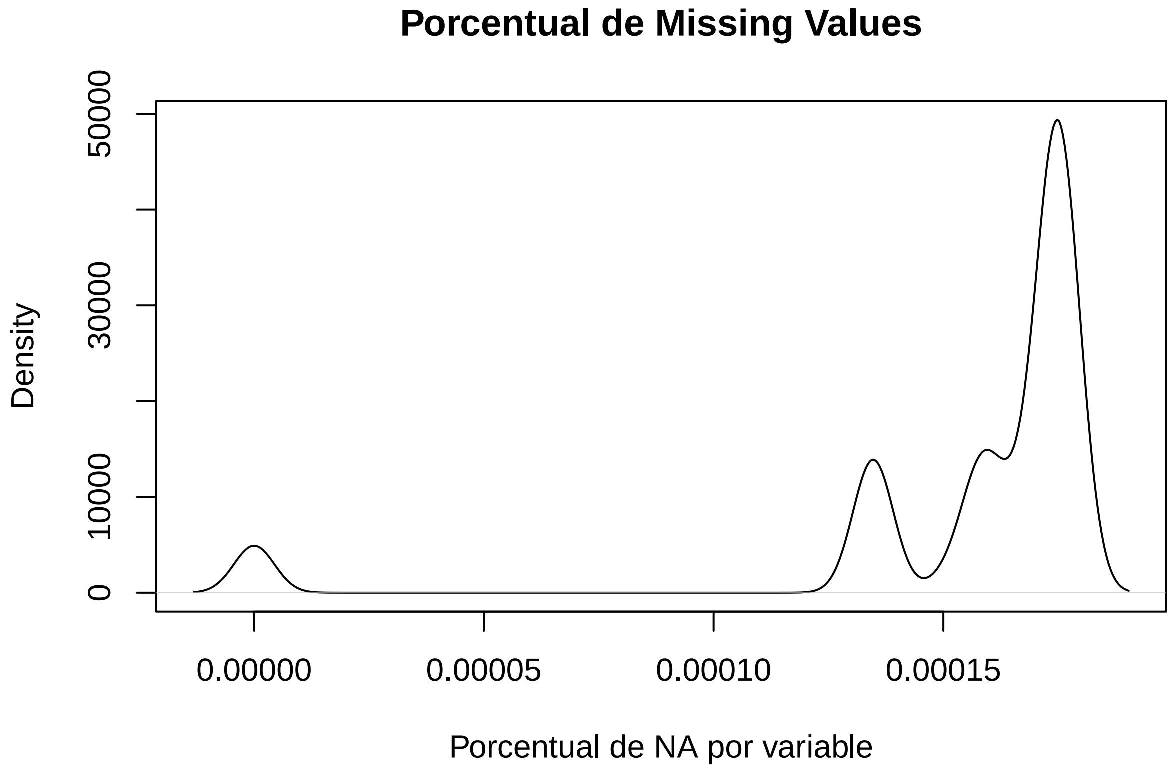
Analisis Exploratorio y Estadistico

La preparacion del dataset se puede encontrar en GitHub, `pre_step_data_einstein.R`:

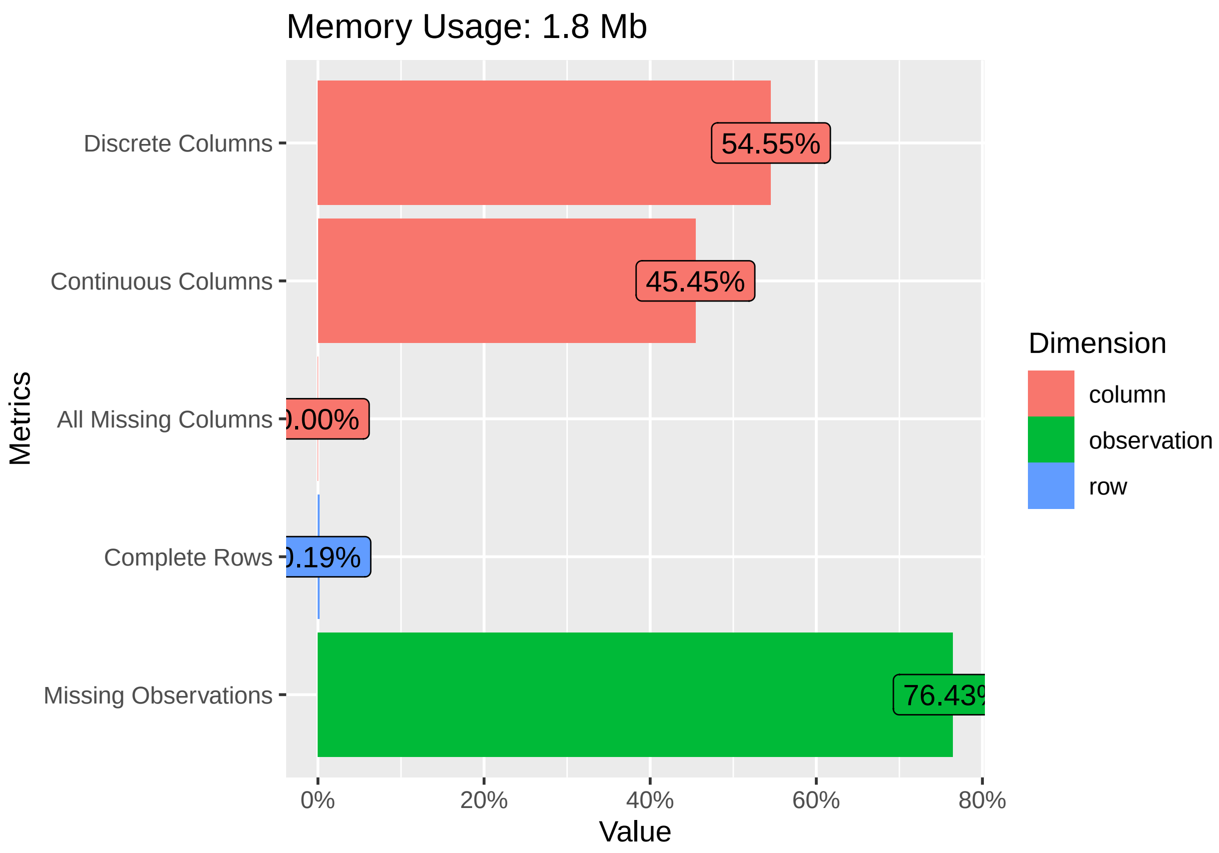
- Las variables se han estandardizados y centrado previamente, por al razon que algunos algortimos funcionan mejor con los datos escalados.
- Se han removido errores, o caracteres especiales, simbolos etc.
- Se han convertido columnas de **character** a **factor, logical** a **factor,** o algunas variables a numerical.
- Se ha creado el **output**, la variable dependiente **care.**



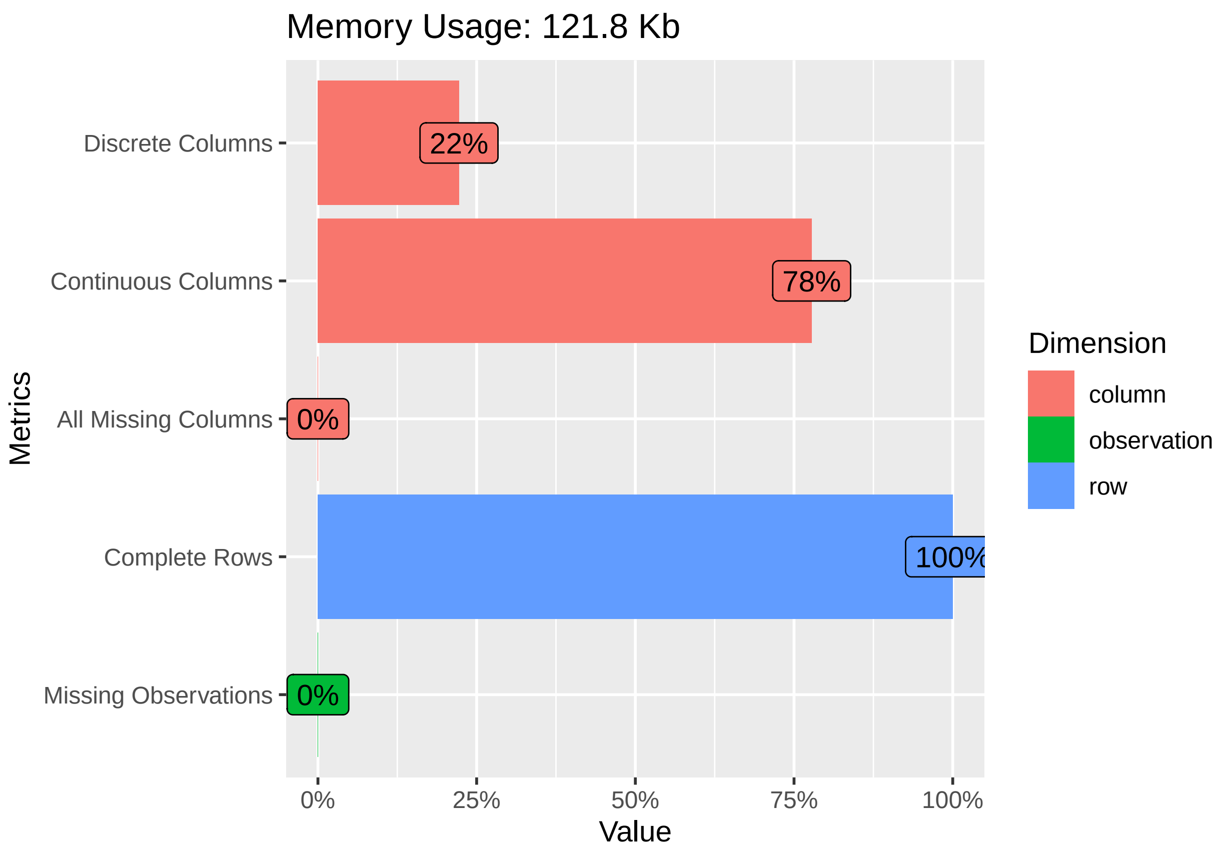
El dataset original, tiene un alto numero de observaciones y variables con **missing values**. Desde el dataset original se han filtrado los datos faltantes en dos maneras.

- En el primer caso se han eliminado pacientes las vcariables que tiene una porcentual de 95% de datos faltantes. Se eliminas tambien las observaciones que tienen por lo menos diez variables con datos.
- En el otro caso se han filtrado las variables con pocos valores, y se han quitado las observaciones que continuaban en tener datos faltantes, de esta forma el dataset se ha notablemente reducido por numero de observaciones y variables. Pero en esto caso no hay datos faltantes.

Como se puede notar en la tabla aqui abajo, todavia hay datos faltantes.



Como se puede ver la segunda opcion es no tener datos faltantes, opcion que se podria tener en cuenta si el numero de observacione fuese suficientemente grande, y sin perder demasiado variables.



Las variables del dataset original estan en el **Anexo - Variables**. Las variables que quedan en los dos dataset, estan resumida en las dos tablas estadisticas en **Anexo 2 - Tablas Estadisticas 1** para el primer caso, y **Anexo 3 - Tablas Estadisticas 2** para el segundo caso.

Primer caso

	negative (N=5086)	positive (N=558)	Total (N=5644)
age_quantile			
- 0	333 (6.5%)	1 (0.2%)	334 (5.9%)
- 1	232 (4.6%)	2 (0.4%)	234 (4.1%)
- 2	310 (6.1%)	5 (0.9%)	315 (5.6%)
- 3	234 (4.6%)	17 (3.0%)	251 (4.4%)
- 4	319 (6.3%)	47 (8.4%)	366 (6.5%)
- 5	250 (4.9%)	44 (7.9%)	294 (5.2%)
- 6	248 (4.9%)	33 (5.9%)	281 (5.0%)
- 7	289 (5.7%)	30 (5.4%)	319 (5.7%)
- 8	140 (2.8%)	27 (4.8%)	167 (3.0%)
- 9	315 (6.2%)	44 (7.9%)	359 (6.4%)
- 10	163 (3.2%)	27 (4.8%)	190 (3.4%)
- 11	340 (6.7%)	40 (7.2%)	380 (6.7%)
- 12	171 (3.4%)	26 (4.7%)	197 (3.5%)
- 13	283 (5.6%)	30 (5.4%)	313 (5.5%)
- 14	260 (5.1%)	39 (7.0%)	299 (5.3%)
- 15	234 (4.6%)	35 (6.3%)	269 (4.8%)
- 16	250 (4.9%)	29 (5.2%)	279 (4.9%)

	negative (N=5086)	positive (N=558)	Total (N=5644)
- 17	244 (4.8%)	19 (3.4%)	263 (4.7%)
- 18	233 (4.6%)	26 (4.7%)	259 (4.6%)
- 19	238 (4.7%)	37 (6.6%)	275 (4.9%)

	discharged (N=5474)	regular_ward (N=79)	semi_intensive (N=50)	intensive_care_unit (N=41)	Total (N=5644)
age_quantile					
- 0	292 (5.3%)	9 (11.4%)	13 (26.0%)	20 (48.8%)	334 (5.9%)
- 1	224 (4.1%)	1 (1.3%)	4 (8.0%)	5 (12.2%)	234 (4.1%)
- 2	311 (5.7%)	3 (3.8%)	1 (2.0%)	0 (0.0%)	315 (5.6%)
- 3	250 (4.6%)	1 (1.3%)	0 (0.0%)	0 (0.0%)	251 (4.4%)
- 4	363 (6.6%)	2 (2.5%)	1 (2.0%)	0 (0.0%)	366 (6.5%)
- 5	292 (5.3%)	2 (2.5%)	0 (0.0%)	0 (0.0%)	294 (5.2%)
- 6	281 (5.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	281 (5.0%)
- 7	315 (5.8%)	3 (3.8%)	1 (2.0%)	0 (0.0%)	319 (5.7%)
- 8	164 (3.0%)	3 (3.8%)	0 (0.0%)	0 (0.0%)	167 (3.0%)
- 9	358 (6.5%)	1 (1.3%)	0 (0.0%)	0 (0.0%)	359 (6.4%)
- 10	186 (3.4%)	3 (3.8%)	1 (2.0%)	0 (0.0%)	190 (3.4%)
- 11	373 (6.8%)	4 (5.1%)	2 (4.0%)	1 (2.4%)	380 (6.7%)
- 12	192 (3.5%)	4 (5.1%)	1 (2.0%)	0 (0.0%)	197 (3.5%)
- 13	304 (5.6%)	6 (7.6%)	3 (6.0%)	0 (0.0%)	313 (5.5%)
- 14	292 (5.3%)	4 (5.1%)	1 (2.0%)	2 (4.9%)	299 (5.3%)
- 15	261 (4.8%)	5 (6.3%)	3 (6.0%)	0 (0.0%)	269 (4.8%)
- 16	275 (5.0%)	3 (3.8%)	0 (0.0%)	1 (2.4%)	279 (4.9%)
- 17	254 (4.6%)	8 (10.1%)	1 (2.0%)	0 (0.0%)	263 (4.7%)
- 18	243 (4.4%)	7 (8.9%)	4 (8.0%)	5 (12.2%)	259 (4.6%)
- 19	244 (4.5%)	10 (12.7%)	14 (28.0%)	7 (17.1%)	275 (4.9%)

Segundo caso

	negative (N=517)	positive (N=81)	Total (N=598)
age_quantile			
- 0	26 (5.0%)	0 (0.0%)	26 (4.3%)
- 1	10 (1.9%)	1 (1.2%)	11 (1.8%)
- 2	18 (3.5%)	1 (1.2%)	19 (3.2%)
- 3	17 (3.3%)	0 (0.0%)	17 (2.8%)
- 4	27 (5.2%)	1 (1.2%)	28 (4.7%)
- 5	15 (2.9%)	5 (6.2%)	20 (3.3%)
- 6	26 (5.0%)	1 (1.2%)	27 (4.5%)
- 7	23 (4.4%)	3 (3.7%)	26 (4.3%)
- 8	12 (2.3%)	2 (2.5%)	14 (2.3%)
- 9	39 (7.5%)	1 (1.2%)	40 (6.7%)
- 10	20 (3.9%)	4 (4.9%)	24 (4.0%)
- 11	37 (7.2%)	5 (6.2%)	42 (7.0%)
- 12	16 (3.1%)	8 (9.9%)	24 (4.0%)
- 13	38 (7.4%)	6 (7.4%)	44 (7.4%)
- 14	26 (5.0%)	9 (11.1%)	35 (5.9%)
- 15	27 (5.2%)	5 (6.2%)	32 (5.4%)
- 16	28 (5.4%)	3 (3.7%)	31 (5.2%)

	negative (N=517)	positive (N=81)	Total (N=598)
- 17	30 (5.8%)	7 (8.6%)	37 (6.2%)
- 18	31 (6.0%)	10 (12.3%)	41 (6.9%)
- 19	51 (9.9%)	9 (11.1%)	60 (10.0%)

	discharged (N=470)	regular_ward (N=57)	semi_intensive (N=42)	intensive_care_unit (N=29)	Total (N=598)
age_quantile					
- 0	9 (1.9%)	1 (1.8%)	6 (14.3%)	10 (34.5%)	26 (4.3%)
- 1	5 (1.1%)	0 (0.0%)	3 (7.1%)	3 (10.3%)	11 (1.8%)
- 2	16 (3.4%)	2 (3.5%)	1 (2.4%)	0 (0.0%)	19 (3.2%)
- 3	16 (3.4%)	1 (1.8%)	0 (0.0%)	0 (0.0%)	17 (2.8%)
- 4	25 (5.3%)	2 (3.5%)	1 (2.4%)	0 (0.0%)	28 (4.7%)
- 5	19 (4.0%)	1 (1.8%)	0 (0.0%)	0 (0.0%)	20 (3.3%)
- 6	27 (5.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	27 (4.5%)
- 7	24 (5.1%)	1 (1.8%)	1 (2.4%)	0 (0.0%)	26 (4.3%)
- 8	13 (2.8%)	1 (1.8%)	0 (0.0%)	0 (0.0%)	14 (2.3%)
- 9	39 (8.3%)	1 (1.8%)	0 (0.0%)	0 (0.0%)	40 (6.7%)
- 10	20 (4.3%)	3 (5.3%)	1 (2.4%)	0 (0.0%)	24 (4.0%)
- 11	37 (7.9%)	2 (3.5%)	2 (4.8%)	1 (3.4%)	42 (7.0%)
- 12	19 (4.0%)	4 (7.0%)	1 (2.4%)	0 (0.0%)	24 (4.0%)
- 13	36 (7.7%)	5 (8.8%)	3 (7.1%)	0 (0.0%)	44 (7.4%)
- 14	28 (6.0%)	4 (7.0%)	1 (2.4%)	2 (6.9%)	35 (5.9%)
- 15	25 (5.3%)	4 (7.0%)	3 (7.1%)	0 (0.0%)	32 (5.4%)
- 16	28 (6.0%)	2 (3.5%)	0 (0.0%)	1 (3.4%)	31 (5.2%)
- 17	29 (6.2%)	7 (12.3%)	1 (2.4%)	0 (0.0%)	37 (6.2%)
- 18	26 (5.5%)	6 (10.5%)	4 (9.5%)	5 (17.2%)	41 (6.9%)
- 19	29 (6.2%)	10 (17.5%)	14 (33.3%)	7 (24.1%)	60 (10.0%)