Islas Canarias
Del 15 al 19 de noviembre de 2021

Encuesta Sanitaria y Social de Hogares: análisis y visualización mediante R y Python

Jorge Hidalgo Calderón
Universidad de Granada
jorgehcal@ugr.es

Luis Castro Martín
Universidad de Granada
luiscastro193@ugr.es

María del Mar Rueda García
Universidad de Granada
mrueda@ugr.es

Andrés Cabrera León
Escuela Andaluza de Salud Pública
andres.cabrera.easp@juntadeandalucia.es

Carmen Sánchez-Cantalejo Garrido
Escuela Andaluza de Salud Pública
carmen.sanchezcantalejo.easp@juntadeandalucia.es

Ramón Ferri García
Universidad de Granada
rferri@ugr.es

Introducción

La Encuesta Sanitaria y SOCial (ESSOC) (Sánchez-Cantalejo et al., 2021) es un proyecto de investigación que surge ante la necesidad de comprender el impacto de la COVID-19 en la población general andaluza y, especialmente, en aquella con mayor riesgo de vulnerabilidad. Sus resultados facilitarán la toma de decisiones en cuanto a medidas de prevención y protección sobre esas poblaciones. La ESSOC recoge información sobre la evolución de características del estado de salud, socioeconómicas, psicosociales, conductuales, laborales, medioambientales y clínicas; tanto de la población general como de colectivos vulnerables. El estudio integra datos de distintas fuentes basadas en encuestas y registros clínicos, epidemiológicos, poblacionales y ambientales. Las encuestas se han realizado mediante un diseño de encuesta panel por superposición.

Este tipo de diseños son ampliamente usados en estudios para observar la evolución de ciertas características a lo largo del tiempo. Sin embargo, implican problemas de no respuesta causados, entre otros factores, por la fatiga de la población al ser encuestada reiteradamente. Este hecho hace necesario llevar a cabo nuevas encuestas hasta completar las muestras transversales de cada medición, reemplazando así a las personas que dejan de contestar de unas mediciones a otras y garantizando de esta manera obtener estimaciones igual de precisas en cada medición.

Después de los ajustes en los pesos muestrales para reducir posibles sesgos debido a la falta de respuesta y de cobertura en este tipo de encuestas panel (Luis Castro et. al), se generan estimaciones (totales, proporciones y de razón) sobre las variables de estudio para las distintas muestras longitudinales y transversales. Estos resultados se almacenan en seis tablas descriptivas para cada variable de estudio, obteniendo por lo tanto una gran cantidad de información que requiere de su visualización en gráficos para facilitar su interpretación.

Objetivos

Mostrar la creación de gráficos interactivos y personalizables para visualizar mediante Python los resultados descriptivos de la ESSOC obtenidos con R

Describir el desarrollo Web donde se almacenarán las tablas descriptivas y sus correspondientes gráficas a modo de repositorio en línea desde donde poder realizar las consultas y descargas correspondientes.

Contexto: tablas descriptivas.

Los resultados del análisis descriptivo de la encuesta ESSOC se recoge en formato de tablas. Los estimadores se dividen en dos tipos: transversales (estimaciones en cada medición y del cambio de una medición con respecto a la primera), o longitudinales (estimaciones sobre la evolución de la diferencia de una medición con respecto a la anterior). En total se obtienen las siguientes seis tablas descriptivas, 4 de ellas a partir de las muestras transversales de cada medición y dos a partir de las muestras longitudinales entre una medición y la anterior:

1. Transversal Variables Originales. Se realizan estimaciones de proporción (porcentajes) sobre todas las categorías de las variables definidas tal cual en el

cuestionario. Se presenta una tabla para cada variable y medición. Además de los porcentajes, se incluye el tamaño muestral, el tamaño poblacional estimado, los intervalos de confianza al 95% y una nota para identificar estimaciones con coeficiente de variación superior al 20%.

		Total					Hombres					Mujeres				
Autopercepción desalud genera		(muestra)	(estimada	n puntual	IC	IC	(muestra	(estimada)	n puntual	IC	IC	n (muestra)	(estimada)	n puntual	IC	IC
Total	Excelente	374	864300	12,4%	11,3%	13,6%	207	480130	15,2%	13,4%	17,2%	167	384170	10,1%	8,7%	11,6%
	Muy buena	676	1569466	22,5%	21,0%	24,1%	329	767612	24,3%	22,1%	26,7%	347	801854	21,0%	19,1%	23,1%
	Buena	1459	3388457	48,6%	46,8%	50,4%	648	1509441	47,8%	45,2%	50,5%	811	1879016	49,2%	46,8%	51,7%
	Regular	423	994467	14,3%	13,0%	15,6%	150	352124	11,2%	9,6%	13,0%	273	642343	16,8%	15,1%	18,7%
	Mala	64	155031	2,2%	1,7%	2,8%	18	45660	1,4%	0,9%	2,3%	46	109371	2,9%	2,1%	3,8%
	Total	2996	6971721	100%			1352	3154967	100%			1644	3816754	100%		
Urbana	Excelente	139	319222	27,4%	23,6%	31,5%	77	176374	36,5%	30,2%	43,3%	62	142848	20,9%	16,6%	26,0%
	Muy buena	170	400055	34,3%	30,2%	38,6%	70	165597	34,3%	28,1%	41,1%	100	234458	34,3%	29,0%	40,0%
	Buena	166	388378	33,3%	29,3%	37,6%	57	131838	27,3%	21,6%	33,8%	109	256540	37,5%	32,1%	43,3%
	Regular	24	58512	5,0%	3,4%	7,4%	4	9073	1,9%	0,7%	4,9%	20	49439	7,2%	4,7%	11,0%
	Mala	0	0	-	-	-	0	0	-	-	-	0	0	-	-	-
	Total	499	1166167	100%			208	482882	100%			291	291	100%		
nsidad interme	Excelente	137	320381	15,9%	13,6%	18,5%	68	161115	17,7%	14,2%	21,9%	69	159266	14,3%	11,4%	17,8%
	Muy buena	239	552881	27,4%	24,5%	30,4%	115	266260	29,3%	25,0%	34,1%	124	286621	25,8%	22,0%	29,9%
	Buena	427	989118	48,9%	45,6%	52,3%	187	433855	47,8%	42,8%	52,8%	240	555263	49,9%	45,4%	54,4%
	Regular	62	144108	7,1%	5,6%	9,1%	19	44098	4,9%	3,1%	7,5%	43	100010	9,0%	6,7%	11,9%
	Mala	6	14441	0,7%	0,3%	1,6%	1	3217	0,4%	0,0%	2,5%	5	11224	1,0%	0,4%	2,4%
	Total	871	2020929	100%			390	908545	100%			481	1112384	100%		
Rural	Excelente	83	190112	6,9%	5,6%	8,5%	54	124092	9,5%	7,3%	12,2%	29	66020	4,6%	3,2%	6,6%
	Muy buena	211	487573	17,8%	15,7%	20,1%	114	266726	20,4%	17,2%	23,9%	97	220847	15,4%	12,8%	18,5%
	Buena	651	1509936	55,1%	52,2%	58,0%	308	718762	54,9%	50,7%	59,0%	343	791174	55,3%	51,4%	59,3%
	Regular	202	473616	17,3%	15,2%	19,6%	75	175467	13,4%	10,8%	16,5%	127	298149	20,9%	17,8%	24,3%
	Mala	32	77954	2,8%	2,0%	4,0%	10	24682	1,9%	1,0%	3,5%	22	53272	3,7%	2,5%	5,6%
	Total	1179	2739191	100%			561	1309729	100%			618	1429462	100%		

2. Transversal Dicotomizada. Se realizan estimaciones de la proporción a partir de variables dicotomizadas según los resultados obtenidos en la tabla anterior, permitiendo de esta manera presentar los porcentajes de las todas las mediciones en la misma tabla. Además de los porcentajes, se incluyen los intervalos de confianza al 95% y una nota para identificar estimaciones con coeficiente de variación superior al 20%.

Autopercepción de salud general excelente o muy buena	Total				Hombres				Mujeres			
	M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)	Estimación porcentual (IC95%)
Total	36,97%	34,98%	35,44%	45,44%	41,74%	37,98%	40,87%	50,87%	32,41%	32,10%	30,23%	40,23%
	(34,68-39,32)	(32,98-37,03)	(33,59-37,33)	(43,59-47,33)	(38,38-45,17)	(35,05-41)	(38,1-43,7)	(48,1-53,7)	(29,33-35,65)	(29,41-34,93)	(27,84-32,73)	(37,84-42,73)
16-29	61,10%	63,26%	62,82%	72,82%	68,83%	64,71%	70,28%	80,28%	53,00%	61,73%	55,01%	65,01%
	(55,41-66,5)	(58,2-68,04)	(58,11-67,31)	(68,11-77,31)	(61,25-75,52)	(57,67-71,17)	(63,52-76,25)	(73,52-86,25)	(44,62-61,23)	(54,34-68,61)	(48,46-61,4)	(58,46-71,4)
30-44	44,53%	42,69%	44,72%	54,72%	47,04%	43,48%	47,98%	57,98%	42,00%	41,90%	41,44%	51,44%
	(39,87-49,28)	(38,68-46,81)	(41,1-48,4)	(51,1-58,4)	(40,21-53,99)	(37,65-49,51)	(42,72-53,29)	(52,72-63,29)	(35,73-48,54)	(36,44-47,57)	(36,54-46,53)	(46,54-56,53)
45-64	29,07%	25,90%	25,00%	35,00%	32,44%	29,44%	29,90%	39,90%	25,77%	22,43%	20,20%	30,20%
	(25,76-32,63)	(23,05-28,97)	(22,49-27,69)	(32,49-37,69)	(27,73-37,53)	(25,3-33,95)	(26,15-33,94)	(36,15-43,94)	(21,23-30,9)	(18,64-26,75)	(17-23,84)	(27-33,84)
65 o más	18,59%	14,48%	16,02%	26,02%	23,06%	17,85%	20,18%	30,18%	15,05%	11,78%	12,70%	22,70%
	(14,41-23,64)	(11,19-18,53)	(12,68-20,02)	(22,68-30,02)	(16,37-31,44)	(12,64-24,6)	(14,61-27,2)	(24,61-37,2)	(10,18-21,69)	(7,91-17,19)	(9,04-17,56)	(19,04-27,56)

3. Transversal Cambio. Se realizan estimaciones de la razón para medir el cambio de los resultados obtenidos para una variable dicotómica en una medición con respecto a la primera. Además de los porcentajes, se incluyen los intervalos de confianza al 95% y una nota para aquellos residuos estandarizados tipificados corregidos superiores a 2 y a 2,5 (útil para identificar asociaciones estadísticamente significativas).

Autopercepcion de salud general excelente o muy buena	Total						Hombres						Mujeres					
	M2 / M1		M3 / M1		M4 / M1		M2 / M1		M3 / M1		M4 / M1		M2 / M1		M3 / M1		M4 / M1	
	cambio	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)
Total	1400	40	100	38	600	30	600	23	200	20	400	15	800	17	600	18	500	15
Almería	300	30	200	30	400	24	100	18	200	15	300	13	200	12	700	15	600	11
Cádiz	500	27	700	29	550	20	400	10	150	12	200	10	100	17	300	19	700	10
Córdoba	600	20	500	24	300	18	400	10	100	7	150	5	200	10	200	17	800	13
Granada	300	30	200	30	400	24	100	18	200	15	300	13	200	12	700	15	600	11
Huelva	500	27	700	29	550	20	400	10	150	12	200	10	100	17	300	19	700	10
Jaén	600	20	500	24	300	18	400	10	100	7	150	5	200	10	200	17	800	13
Málaga	500	27	700	29	550	20	400	10	150	12	200	10	100	17	300	19	700	10
Sevilla	600	20	500	24	300	18	400	10	100	7	150	5	200	10	200	17	800	13

4. Transversal Cambio Brecha de género. Para cada medición con respecto a la primera, se obtiene el estimador de la brecha de género absoluta mediante la diferencia entre los porcentajes de cambio de hombres y de mujeres obtenidos en la tabla anterior (puntos porcentuales). Por otro lado, también se obtiene el estimador de la brecha de género relativo mediante la razón de los cambios de los hombres con respecto a los de las mujeres obtenidos en la tabla anterior. Además de esas estimaciones, se incluyen en las tablas las estimaciones poblacionales, los intervalos de confianza al 95% y una nota para aquellos residuos estandarizados tipificados corregidos superiores a 2 y a 2,5 (útil para identificar asociaciones estadísticamente significativas).

Autopercepcion de salud general excelente o muy buena	Total						Brecha de género absoluta (puntos porcentuales del cambio: Mujeres - Hombres)						Brecha de género relativa (razón porcentual del cambio: Mujeres / Hombres)					
	M2 / M1		M3 / M1		M4 / M1		M2 / M1		M3 / M1		M4 / M1		M2 / M1		M3 / M1		M4 / M1	
	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)	N	(IC95%)
Total	###	40	100	38	600	30	600	23	200	20	400	15	800	17	600	18	500	15
Almería	300	30	200	30	400	24	100	18	200	15	300	13	200	12	700	15	600	11
Cádiz	500	27	700	29	550	20	400	10	150	12	200	10	100	17	300	19	700	10
Córdoba	600	20	500	24	300	18	400	10	100	7	150	5	200	10	200	17	800	13
Granada	300	30	200	30	400	24	100	18	200	15	300	13	200	12	700	15	600	11
Huelva	500	27	700	29	550	20	400	10	150	12	200	10	100	17	300	19	700	10
Jaén	600	20	500	24	300	18	400	10	100	7	150	5	200	10	200	17	800	13
Málaga	500	27	700	29	550	20	400	10	150	12	200	10	100	17	300	19	700	10
Sevilla	600	20	500	24	300	18	400	10	100	7	150	5	200	10	200	17	800	13

5. Longitudinal Diferencia. Para las variables recogidas en dos mediciones consecutivas, se construye una nueva a partir de su diferencia que, posteriormente, se categoriza en ‘Mejor’ o ‘Más’, ‘Igual’, y ‘Peor’ o ‘Menos’. Así pues, para cada variable original, se obtienen tres nuevas variables que miden la diferencia de una medición con respecto a la anterior. Esto solo es posible hacerlo con las muestras longitudinales de al menos dos mediciones consecutivas. Sobre esas nuevas variables, se realiza la estimación porcentual de cada categoría obtenida, así como la estimación poblacional, los intervalos de confianza al 95% y una nota para aquellos residuos estandarizados tipificados corregidos superiores a 2 y a 2,5 (útil para identificar asociaciones estadísticamente significativas).

Autopercepción de salud general		Total						Hombres						Mujeres					
		M2-M1		M3-M2		M4-M3		M2-M1		M3-M2		M4-M3		M2-M1		M3-M2		M4-M3	
		N	Estimación	N	Estimación	N	Estimación	N	Estimación	N	Estimación	N	Estimación	N	Estimación	N	Estimación	N	Estimación
Total	mejora la salud	1449861	20,93% (18,99-23,01)	1441929	20,83% (18,76-23,07)	1441929	20,83% (18,76-23,07)	659426	19,45% (16,82-22,38)	807873	23,85% (20,81-27,18)	807873	23,85% (20,81-27,18)	790435	22,34% (19,59-25,37)	634056	17,94% (15,21-21,03)	634056	17,94% (15,21-21,03)
	permanece igual	3783644	54,61% (52,17-57,03)	3713833	53,65% (51,1-56,18)	3713833	53,65% (51,1-56,18)	1871074	55,18% (51,7-58,62)	1808648	53,39% (49,73-57,02)	1808648	53,39% (49,73-57,02)	1912570	54,06% (50,63-57,45)	1905185	53,89% (50,33-57,41)	1905185	53,89% (50,33-57,41)
	empeora la salud	1694841	24,46% (22,39-26,66)	1767073	25,53% (23,37-27,81)	1767073	25,53% (23,37-27,81)	860057	25,37% (22,42-28,56)	771013	22,76% (19,84-25,97)	771013	22,76% (19,84-25,97)	834784	23,60% (20,76-26,69)	996060	28,17% (25,08-31,49)	996060	28,17% (25,08-31,49)
Urbano	mejora la salud	366732	28,58% (23,63-34,11)	266661	20,79% (16,24-26,22)	266661	20,79% (16,24-26,22)	148078	22,56% (16,8-29,59)	195952	29,86% (22,25-38,79)	195952	29,86% (22,25-38,79)	218654	34,89% (27,16-43,51)	70710	11,29% (7,51-16,64)	70710	11,29% (7,51-16,64)
	permanece igual	603516	47,04% (41,43-52,72)	653950	51,00% (44,72-57,24)	653950	51,00% (44,72-57,24)	324282	49,40% (41,67-57,17)	319098	48,63% (39,75-57,6)	319098	48,63% (39,75-57,6)	279234	44,56% (36,56-52,85)	334852	53,48% (44,68-62,06)	334852	53,48% (44,68-62,06)
	empeora la salud	312800	24,38% (19,75-29,7)	361743	28,21% (22,74-34,41)	361743	28,21% (22,74-34,41)	184058	28,04% (21,59-35,55)	141125	21,51% (14,95-29,93)	141125	21,51% (14,95-29,93)	128742	20,55% (14,3-28,6)	220617	35,23% (27,03-44,41)	220617	35,23% (27,03-44,41)
na intermed	mejora la salud	361696	19,56% (15,91-23,82)	413599	22,39% (18,47-26,87)	413599	22,39% (18,47-26,87)	163167	17,61% (12,63-24,01)	211993	22,89% (17,56-29,27)	211993	22,89% (17,56-29,27)	198529	21,53% (16,46-27,64)	201606	21,88% (16,42-28,55)	201606	21,88% (16,42-28,55)
	permanece igual	1040536	56,28% (51,44-61)	927248	50,19% (45,44-54,94)	927248	50,19% (45,44-54,94)	540103	58,28% (51,2-65,03)	485068	52,37% (45,58-59,08)	485068	52,37% (45,58-59,08)	500433	54,27% (47,61-60,78)	442179	48,00% (41,37-54,7)	442179	48,00% (41,37-54,7)
	empeora la salud	446637	24,16% (20,3-28,48)	506583	27,42% (23,39-31,86)	506583	27,42% (23,39-31,86)	223506	24,12% (18,72-30,49)	229119	24,74% (19,3-31,11)	229119	24,74% (19,3-31,11)	223130	24,20% (18,95-30,35)	277465	30,12% (24,38-36,56)	277465	30,12% (24,38-36,56)
Rural	mejora la salud	457163	19,09% (16,29-22,24)	447632	18,70% (15,83-21,96)	447632	18,70% (15,83-21,96)	231058	19,52% (15,66-24,06)	266509	22,55% (18,47-27,22)	266509	22,55% (18,47-27,22)	226105	18,66% (14,81-23,24)	181123	14,95% (11,14-19,79)	181123	14,95% (11,14-19,79)
	permanece igual	1385263	57,83% (54,03-61,54)	1384497	57,85% (53,95-61,65)	1384497	57,85% (53,95-61,65)	661444	55,87% (50,56-61,05)	659719	55,81% (50,41-61,08)	659719	55,81% (50,41-61,08)	723819	59,75% (54,27-65)	724778	59,84% (54,14-65,28)	724778	59,84% (54,14-65,28)
	empeora la salud	552954	23,08% (20,01-26,47)	561128	23,45% (20,27-26,95)	561128	23,45% (20,27-26,95)	291411	24,61% (20,36-29,42)	255854	21,64% (17,5-26,45)	255854	21,64% (17,5-26,45)	261543	21,59% (17,33-26,56)	305274	25,20% (20,58-30,47)	305274	25,20% (20,58-30,47)

6. Longitudinal Diferencia Brecha de género. Para cada medición con respecto a la anterior, se obtiene el estimador de la brecha de género absoluta mediante la diferencia entre la diferencia de los porcentajes de hombres y de mujeres

obtenidos en la tabla anterior (puntos porcentuales). Por otro lado, también se obtiene el estimador de la brecha de género relativo mediante la razón de los porcentajes de hombres con respecto a los de las mujeres obtenidos en la tabla anterior. Además de esas estimaciones, se incluyen en las tablas las estimaciones poblacionales, los intervalos de confianza al 95% y una nota para aquellos residuos estandarizados tipificados corregidos superiores a 2 y a 2,5 (útil para identificar asociaciones estadísticamente significativas).

Autopercepción de salud general		Total						Brecha de género absoluta: Mujeres - Hombres						Brecha de género relativa : Mujeres / Hombres					
		M2-M1		M3-M2		M4-M3		M2-M1		M3-M2		M4-M3		M2-M1		M3-M2		M4-M3	
		N	Estimación porcentual	N	Estimación porcentual	N	Estimación porcentual	N	Estimació	N	Estimació	N	Estimació	N	Estimación porcentual	N	Estimación porcentual	N	Estimación porcentual
Total	mejora la salud	1E+06	20,93%	####	20,83%	1E+06	20,83%	####	19,45%	####	23,85%	####	23,85%	####	22,34%	####	17,94%	####	17,94%
			(18,99-23,01)		(18,76-23,07)		(18,76-23,07)		(16,82-22,38)		(20,81-27,18)		(20,81-27,18)		(19,59-25,37)		(15,21-21,03)		(15,21-21,03)
	permanece igual	4E+06	54,61%	####	53,65%	4E+06	53,65%	####	55,18%	####	53,39%	####	53,39%	####	54,06%	####	53,89%	####	53,89%
			(52,17-57,03)		(51,1-56,18)		(51,1-56,18)		(51,7-58,62)		(49,73-57,02)		(49,73-57,02)		(50,63-57,45)		(50,33-57,41)		(50,33-57,41)
	empeora la salud	2E+06	24,46%	####	25,53%	2E+06	25,53%	####	25,37%	####	22,76%	####	22,76%	####	23,60%	####	28,17%	####	28,17%
			(22,39-26,66)		(23,37-27,81)		(23,37-27,81)		(22,42-28,56)		(19,84-25,97)		(19,84-25,97)		(20,76-26,69)		(25,08-31,49)		(25,08-31,49)
Urbano	mejora la salud	4E+05	28,58%	####	20,79%	3E+05	20,79%	####	22,56%	####	29,86%	####	29,86%	####	34,89%	####	11,29%	####	11,29%
			(23,63-34,11)		(16,24-26,22)		(16,24-26,22)		(16,8-29,59)		(22,25-38,79)		(22,25-38,79)		(27,16-43,51)		(7,51-16,64)		(7,51-16,64)
	permanece igual	6E+05	47,04%	####	51,00%	7E+05	51,00%	####	49,40%	####	48,63%	####	48,63%	####	44,56%	####	53,48%	####	53,48%
			(41,43-52,72)		(44,72-57,24)		(44,72-57,24)		(41,67-57,17)		(39,75-57,6)		(39,75-57,6)		(36,56-52,85)		(44,68-62,06)		(44,68-62,06)
	empeora la salud	3E+05	24,38%	####	28,21%	4E+05	28,21%	####	28,04%	####	21,51%	####	21,51%	####	20,55%	####	35,23%	####	35,23%
			(19,75-29,7)		(22,74-34,41)		(22,74-34,41)		(21,59-35,55)		(14,95-29,93)		(14,95-29,93)		(14,3-28,6)		(27,03-44,41)		(27,03-44,41)
na intermed	mejora la salud	4E+05	19,56%	####	22,39%	4E+05	22,39%	####	17,61%	####	22,89%	####	22,89%	####	21,53%	####	21,88%	####	21,88%
			(15,91-23,82)		(18,47-26,87)		(18,47-26,87)		(12,63-24,01)		(17,56-29,27)		(17,56-29,27)		(16,46-27,64)		(16,42-28,55)		(16,42-28,55)
	permanece igual	1E+06	56,28%	####	50,19%	9E+05	50,19%	####	58,28%	####	52,37%	####	52,37%	####	54,27%	####	48,00%	####	48,00%
			(51,44-61)		(45,44-54,94)		(45,44-54,94)		(51,2-65,03)		(45,58-59,08)		(45,58-59,08)		(47,61-60,78)		(41,37-54,7)		(41,37-54,7)
	empeora la salud	4E+05	24,16%	####	27,42%	5E+05	27,42%	####	24,12%	####	24,74%	####	24,74%	####	24,20%	####	30,12%	####	30,12%
			(20,3-28,48)		(23,39-31,86)		(23,39-31,86)		(18,72-30,49)		(19,3-31,11)		(19,3-31,11)		(18,95-30,35)		(24,38-36,56)		(24,38-36,56)
Rural	mejora la salud	5E+05	19,09%	####	18,70%	4E+05	18,70%	####	19,52%	####	22,55%	####	22,55%	####	18,66%	####	14,95%	####	14,95%
			(16,29-22,24)		(15,83-21,96)		(15,83-21,96)		(15,66-24,06)		(18,47-27,22)		(18,47-27,22)		(14,81-23,24)		(11,14-19,79)		(11,14-19,79)
	permanece igual	1E+06	57,83%	####	57,85%	1E+06	57,85%	####	55,87%	####	55,81%	####	55,81%	####	59,75%	####	59,84%	####	59,84%
			(54,03-61,54)		(53,95-61,65)		(53,95-61,65)		(50,56-61,05)		(50,41-61,08)		(50,41-61,08)		(54,27-65)		(54,14-65,28)		(54,14-65,28)
	empeora la salud	6E+05	23,08%	####	23,45%	6E+05	23,45%	####	24,61%	####	21,64%	####	21,64%	####	21,59%	####	25,20%	####	25,20%
			(20,01-26,47)		(20,27-26,95)		(20,27-26,95)		(20,36-29,42)		(17,5-26,45)		(17,5-26,45)		(17,33-26,56)		(20,58-30,47)		(20,58-30,47)

A su vez, en cada una de las tablas los resultados son segmentados por Sexo y por otra variable de estratificación a elegir entre Grupos de Edad, Grado de Urbanización o Provincia. De esta manera, estas tablas proporcionan resultados descriptivos exhaustivos y segmentados (según sexo, grupos de edad, grado de urbanización y provincia) sobre las variables recogidas en las diferentes mediciones de la ESSOC, tanto para cada medición de manera independiente como para los cambios y diferencias de unas mediciones a otras. Además, los coeficientes de variación permiten detectar estimaciones poco precisas mientras que los intervalos de confianza y los residuos permiten identificar asociaciones estadísticamente significativas.

Resultados: tablas descriptivas

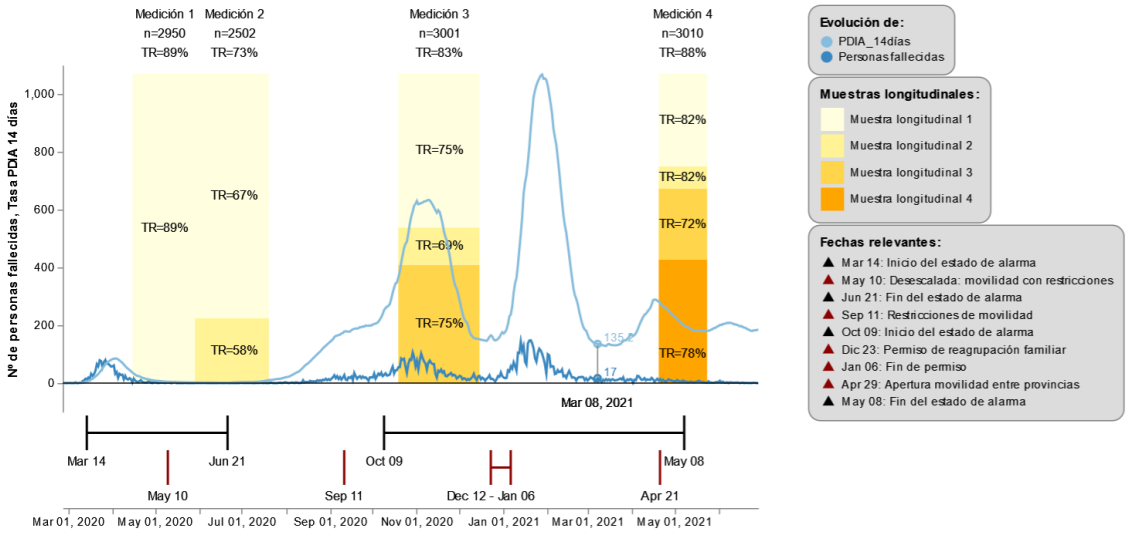
La ESSOC recoge más de 1000 variables originales que, sumadas a las variables nuevas dicotomizadas, proporcionan miles de tablas con decenas de resultados en su interior. Esta cantidad ingente de información provoca la pregunta de ¿qué podemos hacer para digerirla, identificar la información relevante y poder así plantear hipótesis y/o conclusiones?

La respuesta más inmediata es mediante la visualización de estos datos, es decir, la creación de gráficas que ilustren las estimaciones y sus intervalos de confianza, las evoluciones temporales y comparaciones entre mediciones y la personalización mediante la selección de los estadísticos y variables de segmentación, entre otras funcionalidades.

Contexto: utilización de Altair, Python

Las visualizaciones de las tablas descriptivas se han creado con Altair, Python (https://altair-viz.github.io/). Esta biblioteca es la que hemos encontrado más adecuada para esta tarea ya que permite un alto nivel de personalización, de adjuntar gráficos, y de añadir interactividad. Por estas capacidades la hemos elegido frente a bibliotecas más típicas como MatPlotlip, SeaBorn o Plotnine.

Un ejemplo de un gráfico creado para mostrar la evolución de la pandemia por COVID-19 y del trabajo de campo de la ESSOC es el siguiente:



En este gráfico se aprecia de manera clara la tasa PDIA (Pruebas Diagnósticas de Infección Activa), el número de fallecidos, tasas de respuesta de las muestras longitudinales y transversales, el número de personas encuestadas en cada medición, las fechas del trabajo de campo, y varias fechas claves relacionadas con el Covid-19. Además, al pasar el ratón por la gráfica de muestran los valores de las curvas azules en un momento dado.

Metodología: procesamiento de las tablas descriptivas con Python.

Al leer las tablas, es necesario un breve procesamiento de los datos para poder crear las visualizaciones. Esta tarea se ha desarrollado de manera estándar usando la librería Pandas de Python que es la más común para el tratamiento de datos. Los principales problemas de este procesamiento en particular son:

1. Trabajar con la variable Sexo y las tres variables de estratificación (Grupos de Edad, Provincia, Grado de Urbanización), es decir, con tres tablas, a la vez, al mismo tiempo que se trabaja con las estimaciones puntual y de intervalos de confianza.
2. Las variables de estratificación cruzadas con el Sexo complican escribir directamente el código que genere las gráficas, así como la posición de los intervalos de confianza de las estimaciones en las tablas.

El primer problema tiene la siguiente solución: crear un objeto donde se puedan referenciar todas las distintas tablas a la vez para trabajar conjuntamente. En este caso, simplemente he usado una lista de dataframes. Así, por ejemplo

$$df[l][i]\ para\ l\ \in [0,1,2], i\ \in\ [0,1],$$

es estimación porcentual para $i\ =\ 0$, total para $i\ =\ 1$, y la variable Edad para $l\ =\ 0$, Provincia para $l\ =\ 1$, y Grado de Urbanización para $l\ =\ 2$. De esta forma, bucles anidados permiten trabajar con todas estas variantes de dataframes a la vez.

Respecto al segundo problema, lo más cómodo para resolverlo es desagregar las variables de estratificación y hacer de la medición una variable. Para ello se ha usado el siguiente código (se muestra un ejemplo que es análogo para el resto).

Dataframe inicial:

```
In [20]: df[0][0]
```

Out[20]:

	autop	total_M2-M1_Porc	total_M3-M2_Porc	total_M4-M3_Porc	hombres_M2-M1_Porc	hombres_M3-M2_Porc	hombres_M4-M3_Porc	mujeres_M2-M1_Porc	mujeres_M3-M2_Porc	mujeres_M4-M3_Porc	edad
0	mejora la salud	0.209265	0.208286	0.208286	0.194489	0.238484	0.238484	0.223426	0.17935	0.17935	Total
1	NaN	(18,99-23,01)	(18,76-23,07)	(18,76-23,07)	(16,82-22,38)	(20,81-27,18)	(20,81-27,18)	(19,59-25,37)	(15,21-21,03)	(15,21-21,03)	Total
2	permanece igual	0.546111	0.536461	0.536461	0.551849	0.533913	0.533913	0.540612	0.538903	0.538903	Total
3	NaN	(52,17-57,03)	(51,1-56,18)	(51,1-56,18)	(51,7-58,62)	(49,73-57,02)	(49,73-57,02)	(50,63-57,45)	(50,33-57,41)	(50,33-57,41)	Total
4	empeora la salud	0.244624	0.255253	0.255253	0.253662	0.227603	0.227603	0.235962	0.281747	0.281747	Total
5	NaN	(22,39-26,66)	(23,37-27,81)	(23,37-27,81)	(22,42-28,56)	(19,84-25,97)	(19,84-25,97)	(20,76-26,69)	(25,08-31,49)	(25,08-31,49)	Total
6	mejora la salud	0.285829	0.207947	0.207947	0.225585	0.298627	0.298627	0.348936	0.112922	0.112922	16-29
7	NaN	(23,63-34,11)	(16,24-26,22)	(16,24-26,22)	(16,8-29,59)	(22,25-38,79)	(22,25-38,79)	(27,16-43,51)	(7,51-16,64)	(7,51-16,64)	16-29

Dataframe final:

```
In [22]: df[0][0]
```

Out[22]:

	edad	autop	Medida	est	lim_inf	lim_sup
0	Total	mejora la salud	total_M2-M1_Porc	20.9265	18.99	23.01
1	Total	permanece igual	total_M2-M1_Porc	54.6111	52.17	57.03
2	Total	empeora la salud	total_M2-M1_Porc	24.4624	22.39	26.66
3	16-29	mejora la salud	total_M2-M1_Porc	28.5829	23.63	34.11
4	16-29	permanece igual	total_M2-M1_Porc	47.0377	41.43	52.72
...	...	...	...	...	...	...
130	45-65	permanece igual	mujeres_M4-M3_Porc	59.8409	54.14	65.28
131	45-65	empeora la salud	mujeres_M4-M3_Porc	25.2048	20.58	30.47
132	65+	mejora la salud	mujeres_M4-M3_Porc	23.2545	16.34	31.97
133	65+	permanece igual	mujeres_M4-M3_Porc	51.9349	43.51	60.25
134	65+	empeora la salud	mujeres_M4-M3_Porc	24.8106	18.45	32.49

135 rows × 6 columns

Lo que se consigue es organizar los datos de forma que cada fila corresponde a un individuo y juntar en la misma fila la estimación y los límites del intervalo de confianza. El código para hacer este cambio se basa en el atributo *.melt* de los dataframes en panda:

```
for l in range(0,3):
   lim_inf = []
   lim_sup = []
   for i in range(0,2):
      if i == 0:
         for col in df[l][i].columns:
            if col != var_est[l] and col != 'autop':
               for k in range(0, len(df[l][i])):
                  if k % 2 != 0:
                     lim_inf.append(float(df[l][i][col][k].split('-')[0].replace(',', '.').replace('(', '')))
                     lim_sup.append(float(df[l][i][col][k].split('-')[1].replace(',', '.').replace(')', '')))
         df[l][i] = df[l][i].dropna()
         df[l][i] = df[l][i].melt(id_vars=[var_est[l], 'autop'], var_name='Medida',value_name='est')
         df[l][i]['est'] = df[l][i]['est'] * 100
         df[l][i]['lim_inf'] = lim_inf
         df[l][i]['lim_sup'] = lim_sup
      else:
         df[l][i] = df[l][i].dropna()
         df[l][i].index = range(0, len(df[l][i]))
         df[l][i] = df[l][i].melt(id_vars=[var_est[l], 'autop'], var_name='Medida',value_name='est')
```

Este código almacena los valores de lim_inf y lim_sup, desagrega las variables con el atributo *.melt*, y luego pega la información de los intervalos en el orden adecuado.

Metodología: visualización con Altair (Python)

Una vez que se tiene el dataframe bien organizado, se procede a crear los gráficos usando Altair. La idea es almacenar estos gráficos que se van creando para posteriormente mostrarlos en la Web de la ESSOC, por lo que el código siguiente muestra cómo se rellena la lista graphics[l][i] siguiendo la misma notación que en los dataframes:

```
graphics = []
subgraphics = []

for l in range(0,3):
   for i in range(0, 2):
      bar_chart = alt.Chart(df[l][i]).mark_bar(
      ).encode(
         x=alt.X('Medición:N', title='Medición', axis=alt.Axis(labelAngle=0)),
         y=alt.Y('est:Q', title=tit[i%2]),
         order='order',  # this controls stack order
         color=alt.Color('autop:N', scale=alt.Scale(scheme='spectral', reverse=True),
                     sort=alt.EncodingSortField('order', order='descending'),
                     legend=alt.Legend(title="",
                                 symbolSize=400, symbolType='square', labelLimit=0)),
         tooltip=t[i%2]
      ).add_selection(
         selection
      ).transform_filter(
         selection
      ).properties(
         width=anchura[l],
         height=180
      )

      bar_chart = bar_chart.facet(
      row=alt.Row('Sexo:N', title=None, header=alt.Header(labelFontSize=15)),
      column=alt.Column(var_est[l], title=None, header=alt.Header(labelFontSize=15), sort =
sorts[l])).interactive()




       leyenda = alt.Chart(df[l][i]).mark_text().encode(
          color=alt.Color('autop:N', scale=alt.Scale(scheme='spectral', reverse=True),
                     sort=alt.EncodingSortField('order', order='descending'),
                     legend=alt.Legend(title="",
                                 symbolSize=400, symbolType='square', labelLimit=0, offset= 0)),
          opacity=alt.value(0.2)
       ).add_selection(selection
                   )

       trans_orig = alt.hconcat(bar_chart, leyenda,  title='Autopercepción de salud general en
 cada medición según sexo y '+ var_est[l]  +' (' + tipo[i] + ')'
          ).configure_view(strokeWidth=0).configure_title(
          color='black',
          dy=-20,
          dx=70,
          fontSize=15)

       subgraphics.append(trans_orig)
    graphics.append(subgraphics)
    subgraphics = []
```

Lo que se genera son dos gráficas que se juntan para cada l, i, la principal y la leyenda. Altair ha sido particularmente útil para esto ya que la interactividad específica que buscábamos no la hemos encontrado con otro lenguaje de programación ni programa. Lo crucial aquí es que la leyenda es interactiva y permite cambiar la información que se presenta en la gráfica principal manteniendo la estructura de esta intacta.

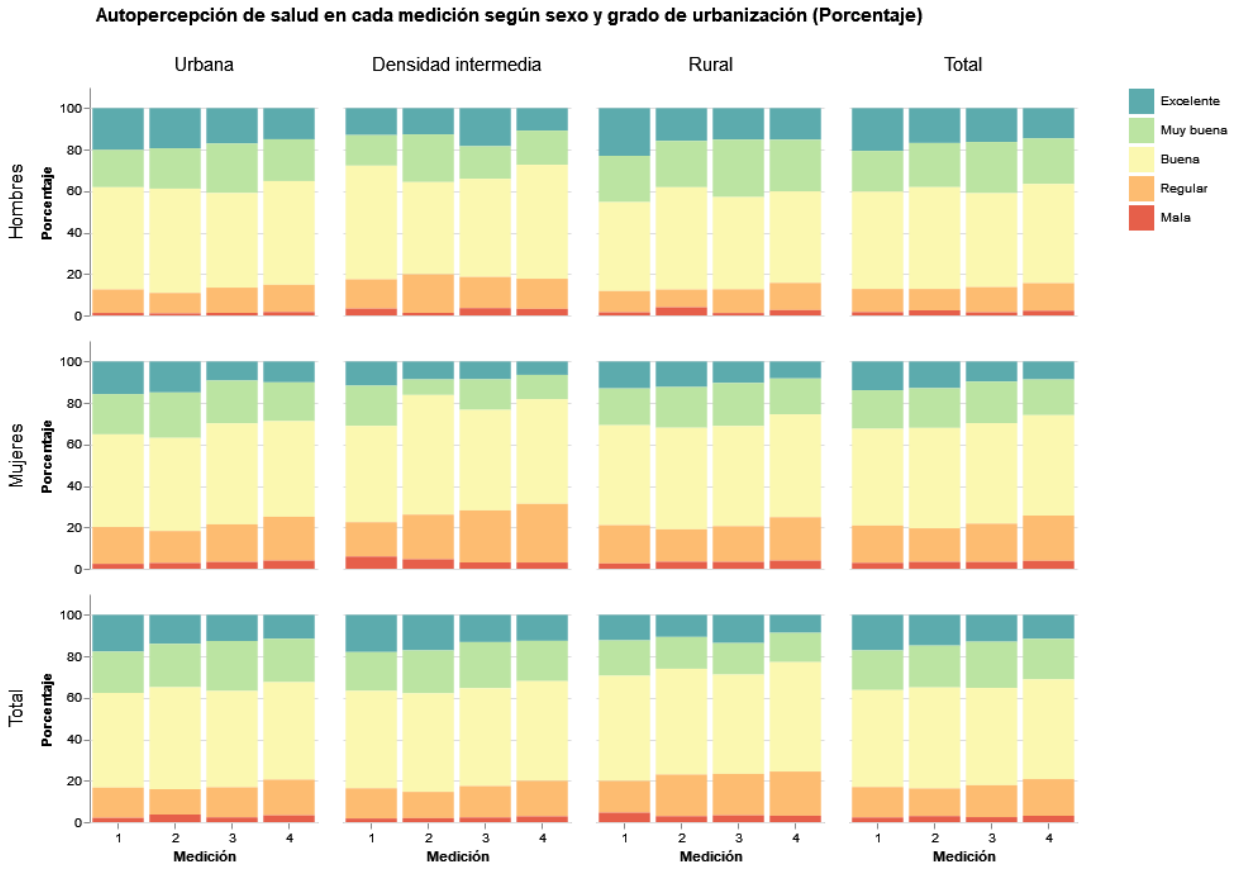
Todo esto se consigue usando los tooltips y los selectores de la librería Altair, y concatenando las dos gráficas (en lugar de simplemente añadir la leyenda por defecto), para que así queden vinculadas las acciones que se hacen en la leyenda con lo que se muestra en el gráfico principal.

Resultados: ejemplos de gráficas

Para la variable: Autopercepción de salud. Más específicamente, esta variable se corresponde con la pregunta: En general diría que su salud es…

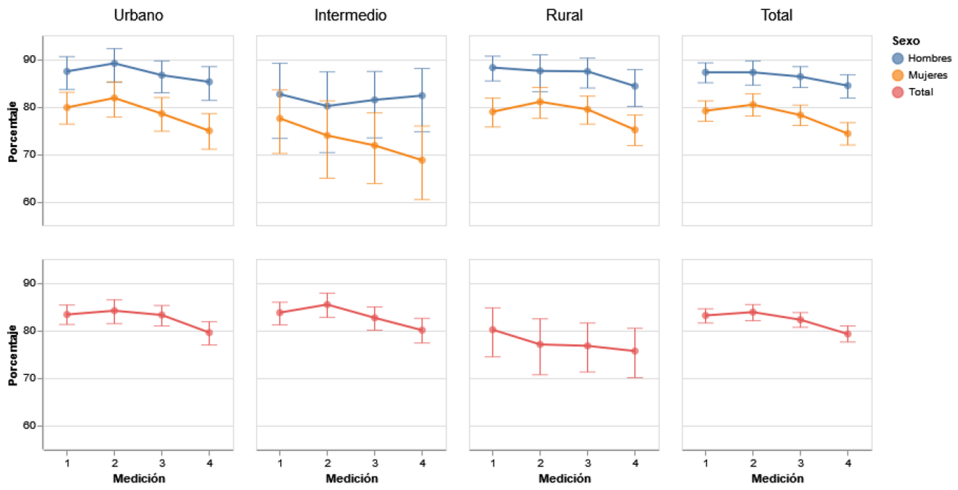
1. Excelente
2. Muy Buena
3. Buena
4. Regular
5. Mala

Transversal Variables Originales: Autopercecpción de salud en cada medición según sexo y grado de urbanización (Porcentaje)

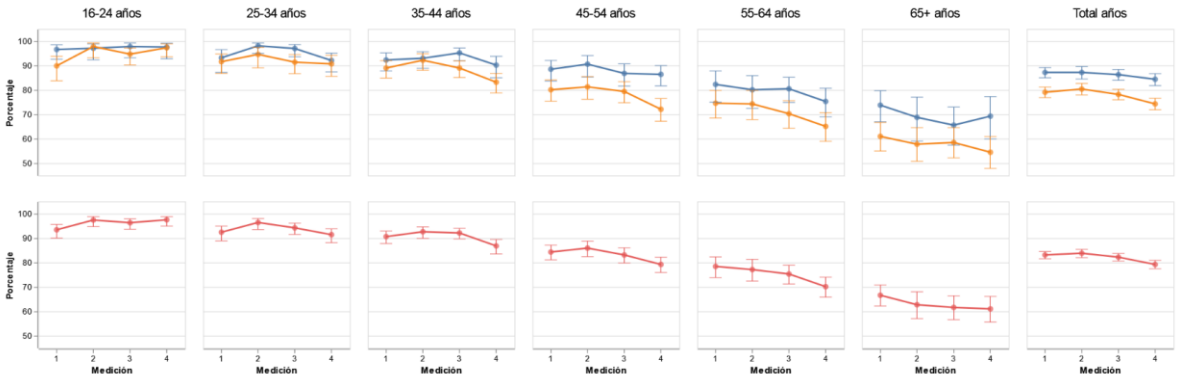


Se observa cómo predomina la percepción de salud Buena, que se corresponde a la respuesta neutra. La salud Excelente y Muy Buena predomina sobre la Regular y Mala, aunque sin mucha diferencia, aunque se observa cómo el primer grupo va perdiendo su dominio gradualmente para cedérselo al segundo. Esto se observa mejor en el siguiente gráfico.

Transversal Variables Dicotómicas: Autopercepción de salud general Excelente o Muy Buena en cada medición según sexo y grado de urbanización.



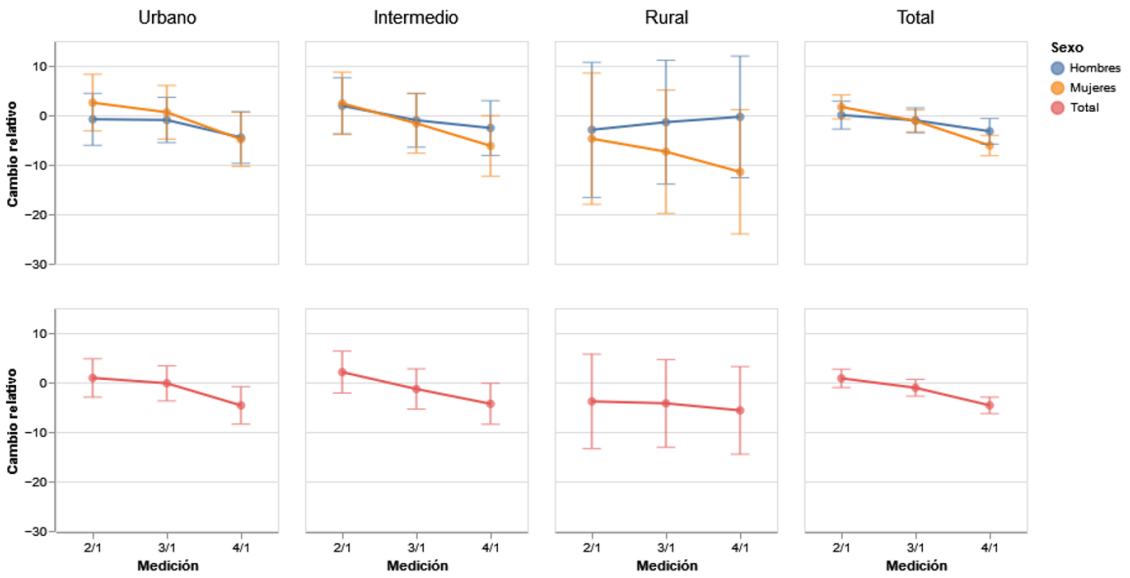
Autopercepción de salud general Excelente o Muy Buena en cada medición según sexo y edad.



.

Se observa claramente que la autopercepción de salud excelente va decreciendo, es decir, la población se siente peor conforme pasan las mediciones, siendo las mujeres en general las que se auto-perciben con peor salud. Nótese que hay una clara diferencia entre hombres y mujeres en este caso.

Transversal Cambio: Cambio relativo en la autopercepción de salud Excelente o Muy buena en cada medición según sexo y grado de urbanización.

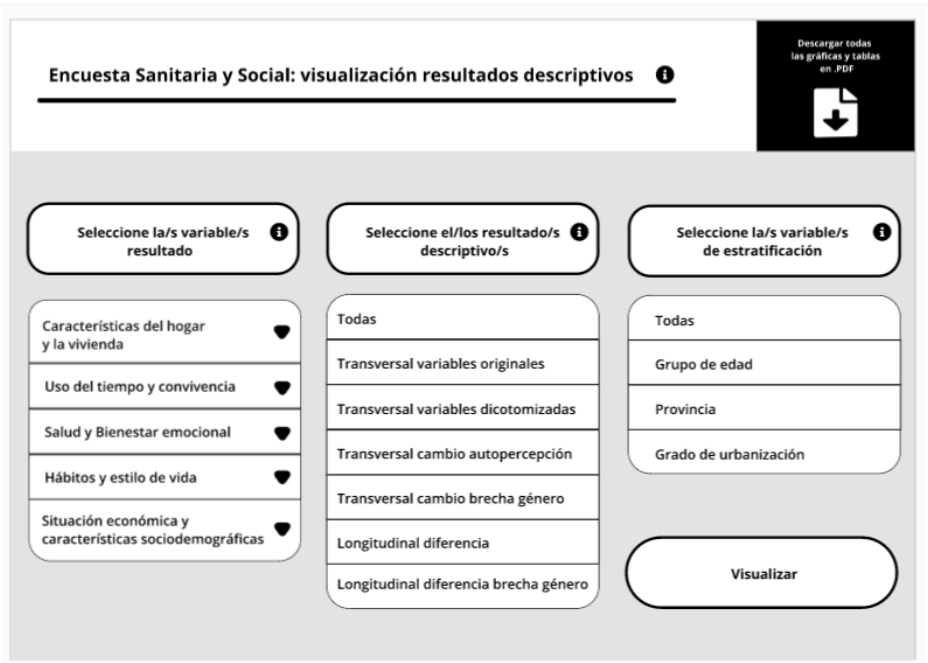


Se observa claramente que la autopercepción de salud excelente respecto de la primera medición va cambiando negativamente. No se aprecia ninguna diferencia significativa entre hombres y mujeres en este caso.

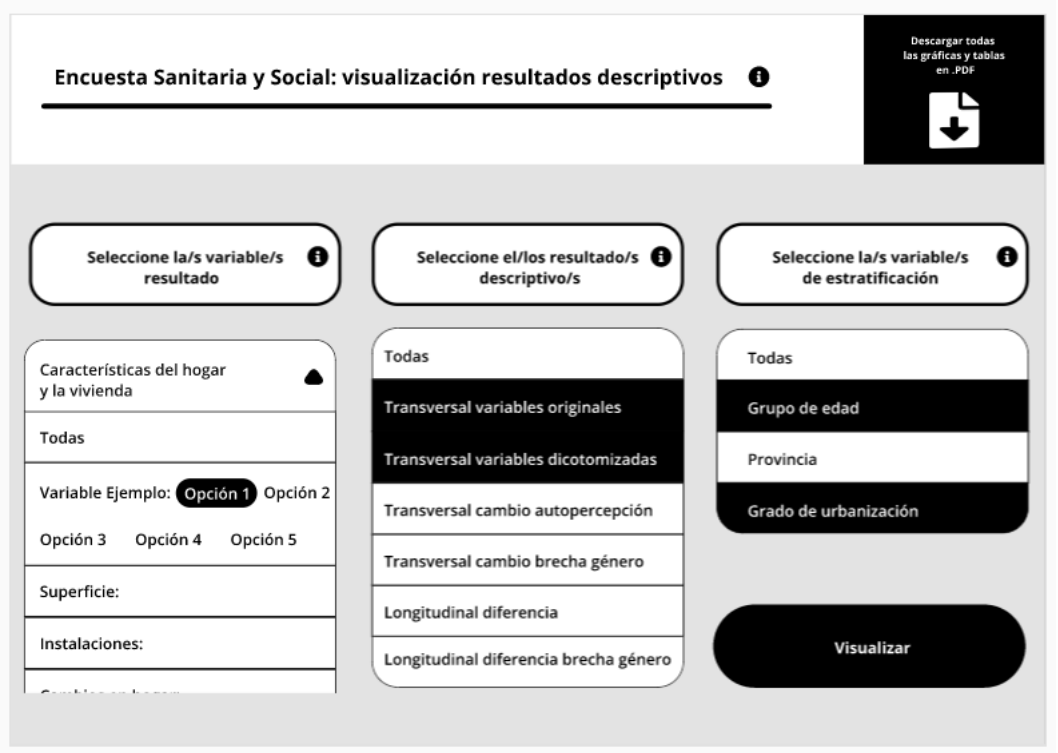
Resultados: Web

La Web de la ESSOC se encuentra actualmente en desarrollo, pero la estructura y el diseño están concretados mediante el software Adobe XD. Esta Web tiene como objetivo servir de repositorio online interactivo para proporcionar de manera cómoda los resultados del análisis de la ESSOC a los usuarios.

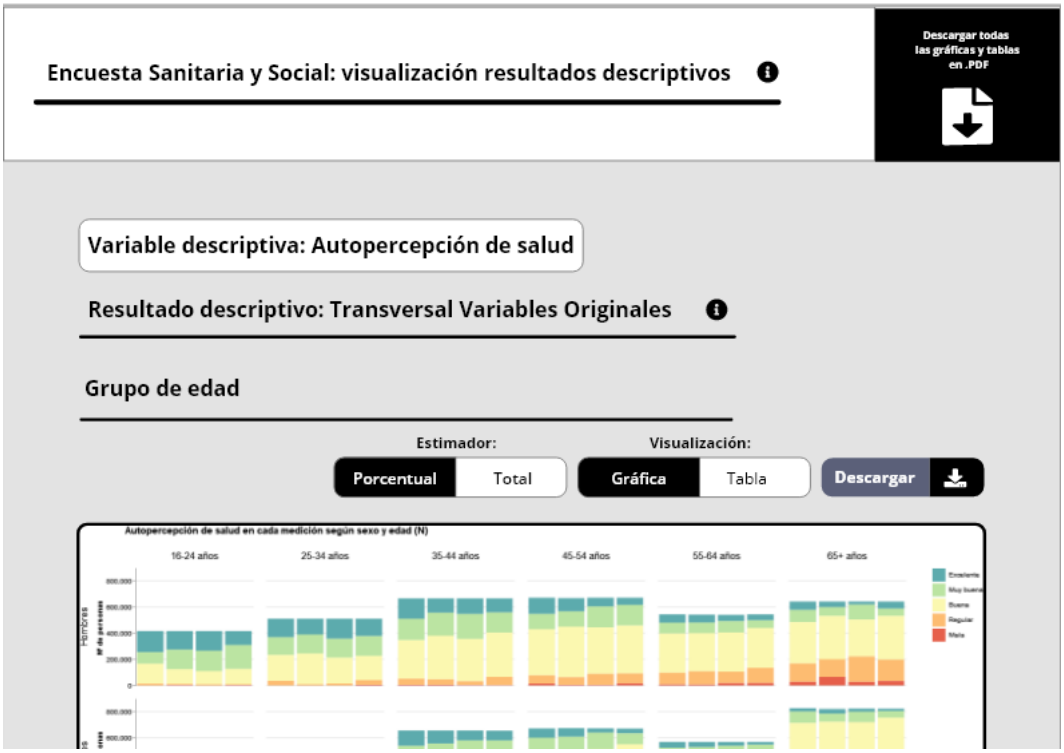
La página principal de la Web:



Se pueden descargar todas las tablas y gráficas del proyecto pulsando arriba a la derecha y elegir una o varias variables resultados, uno o varios resultados descriptivos y una o varias variables de estratificación:



para después pulsar Visualizar e ir a otra ventana donde se muestran los resultados elegidos en orden:



La ventaja de usar los selectores y venir a esta pantalla es que se tiene un alto nivel de interacción para mostrar información concreta en pantalla, y que se puede descargar o bien la selección conjunta de interés, o bien las tablas y gráficas de interés individualmente. Además, las gráficas se muestran embebidas en formato html y se puede interactuar con ellas.

Metodología: Automatización del proceso

Una vez que se ha realizado el proceso de generar las gráficas para una variable, este se debe generalizar para poder manejar las más de 1000 variables que hay en la encuesta. Claramente, hacer esto a mano es inviable, por lo que se crea un diccionario en Python donde referenciando la variable se obtiene información particular de esta que permite crear su correspondiente gráfica. Dicho de otra manera, el diccionario permite crear una función general que tome como parámetro la variable y que genere sus visualizaciones. Un ejemplo para unas pocas variables luce como lo siguiente:

```
maintitle = {
 '16': 'Autopercepción de salud',
 '17': 'Autopercepción de salud mental',
 '18_1': 'Sentimiento de depresión',
 '18_3': 'Sentimiento de soledad',
 '24_5':'Dolor de cabeza',
 '24_6': 'Dolores musculares o de articulaciones',
 '24_7':'Otros dolores'
}
```

Así, referenciando a maintitle[‘16’], se obtiene uno de los títulos de las gráficas de la variable 16 (que corresponde a Autopercepción de salud general). Esto se generaliza y se crea un diccionario de diccionarios que almacenan toda la información como títulos, números de filas y columnas que se deben leer en el Excel de donde se extrae la información, el orden en que se apilan las categorías en el caso de Transversal Originales, etc.

Conclusión

La visualización de datos es esencial para un buen entendimiento de estos, y para facilitar la tarea de desarrollar hipótesis y conclusiones a aquellas personas que trabajen con estos datos. En nuestro caso, la biblioteca Altair ha resultado de extrema utilidad para desarrollar todas las gráficas que queríamos y la recomendamos a aquellas personas que necesiten en algún momento crear visualizaciones de datos.

Financiación

Se ha obtenido financiación de las convocatorias competitivas del Fondo SUPERA COVID-19 de Santander Universidades (SAUN), la Conferencia de Rectores de Universidades Españolas (CRUE), y el Consejo Superior de Investigaciones Científicas (CSIC), además del Programa de Ayudas Competitivas COVID-19 de Pfizer Global Medical Grants. Este trabajo también está apoyado en parte por la beca IMAG-Maria de Maeztu CEX2020-001105-M/AEI/10.13039/501100011033 y por el Ministerio de Ciencia e Innovación, España [número de beca PID2019-106861RB-I00/AEI/10.13039/501100011033].

Referencias

- Sánchez-Cantalejo, C., Rueda, M. M., Saez, M., Enrique, I., Ferri-García, R., De la Fuente, M., Castro-Martín, L., ... & Cabrera-León, A. (2021). Impact of COVID-19 on the health of the general and more vulnerable population and its determinants: Health care and social survey-ESSOC, study protocol.
- Luis Castro-Martín, María del Mar Rueda, Andrés Cabrera, Carmen Sánchez-Cantalejo, Ramón Ferri-García, Jorge Hidalgo. Reweighting with machine learning techniques in panel surveys. Application to the Health Care and Social Survey. The 19th Conference of the Applied Stochastic Models and Data Analysis International Society ASMDA2021
- https://altair-viz.github.io/