



Instituto Tecnológico
de Buenos Aires

82.04 - Analitica Descriptiva

Trabajo Práctico Final

Segundo cuatrimestre de 2021

Collado, Camila (61487)

Noguera, Abril (61541)

Pettinato, Camila (61050)

SCRIPT

Introducción

Para las finalidades del Trabajo Práctico Final se seleccionó la base de datos de la encuesta permanente de hogares de Argentina (EPH), es un programa nacional de producción sistemática y permanente de indicadores sociales que lleva a cabo el INDEC, que permite conocer las características sociodemográficas y socioeconómicas de la población.

A partir de ella nos resultó interesante estudiar los efectos que trajo la pandemia, no tanto en el sentido de salud, si no en la organización de planes y ayudas que se produjeron a partir de ella. Es decir, a partir de la pandemia y las medidas económicas qué impuso el gobierno para ayudar a afrontar la crisis cómo cambiaron los panoramas en tres periodos distintos: pre pandemia (1er trimestre del 2020), en plena pandemia (3er trimestre del 2020) y en un sentido post pandemia, cuando se relajó un poco todo (1er trimestre del 2021).

Para dar contexto, investigamos cuáles fueron las ayudas económicas ofrecidas por el gobierno durante la pandemia y seleccionamos aquellas qué nos resultaron interesantes ver el efecto en el dataset. Las ayudas económicas elegidas son:

- Ingreso familiar de emergencia (IFE) es un bono de \$10.000 que busca paliar el impacto de la emergencia sanitaria sobre la economía de las familias argentinas más afectadas. Para poder obtenerlo debe ser argentino nativo o naturalizado, y haber vivido un mínimo de 2 años en el país, tener entre 18 y 65 años y no tener otros ingresos (trabajos en relación de dependencia, monotributos de categoría C o superior, o del régimen de autónomos, prestaciones de desempleo, jubilaciones, pensiones o retiros contributivos o no contributivos nacionales, provinciales, municipales o de la Ciudad Autónoma de Buenos Aires, planes sociales, salario social complementario, Hacemos Futuro, Potenciar Trabajo u otros programas sociales nacionales, provinciales o municipales)
- Congelamiento de alquileres y suspensión de desalojos
- Suspensión temporaria del corte de servicios por falta de pago
- Suspensión del cierre de cuentas bancarias.
- Suspensión de las comisiones por extracción en cajeros automáticos.

Lo que se quiere analizar es si estas reformas funcionaron para sostener la economía, si la crisis post pandemia afectó a la sociedad y si el gobierno logró mantener la economía.

Análisis descriptivo de bases de datos:

Como se mencionó anteriormente, nuestro análisis será a partir de las distintas medidas económicas que impuso el gobierno para mantener la estabilidad económica del país. Las preguntas que motivaron para la selección de la base fueron:

- ¿Hubo un efecto en la variación del monto de plan según las regiones? por género? por edad?
- ¿Cómo evolucionan en el tiempo las variables económicas?
- ¿Es posible estimar la evolución de estas variables en el futuro?
- ¿Hay relación entre la cantidad de planes y el ingreso familiar?

Para comenzar con el análisis debemos aclarar que las conclusiones se calcularán a partir de una muestra. Y como una persona entrevistada puede no responder toda la encuesta decidimos filtrar aquellas qué sí o sí respondieron la entrevista individual para estar seguros de qué esa parte cómo mínimo la tienen todos. Para eso, se utiliza el campo H15: Entrevista individual realizada.

Se descartaron varias variables debido a que la base presenta demasiadas y que probablemente no aporten al análisis. Las variables seleccionadas fueron:

- CODUSU: Código para distinguir VIVIENDAS, permite aparearlas con Hogares y Personas. Además permite hacer el seguimiento a través de los trimestres.
- NRO_HOGAR: Código para distinguir HOGARES.
- COMPONENTE: Nº de orden que se asigna a las personas que conforman cada hogar de la vivienda.
- ANO4: Año de relevamiento
- TRIMESTRE
- REGION: Código de Región
 - 01 = Gran Buenos Aires
 - 40 = Noroeste
 - 41 = Nordeste
 - 42 = Cuyo
 - 43 = Pampeana
 - 44 = Patagónica
- AGLOMERADO
- PONDERA
- CH04: Sexo.
- CH06: Edad.
- CH10: ¿Asiste o asistió a algún establecimiento educativo?(colegio, escuela, universidad)
- NIVEL_ED: NIVEL EDUCATIVO

- 1 = Primaria Incompleta(incluye educación especial)
 - 2 = Primaria Completa
 - 3 = Secundaria Incompleta
 - 4 = Secundaria Completa
 - 5 = Superior Universitaria Incompleta
 - 6 = Superior Universitaria Completa
 - 7 = Sin instrucción
 - 9 = Ns./ Nr.
- ESTADO: CONDICIÓN DE ACTIVIDAD
 - 1 = Ocupado
 - 2 = Desocupado
 - 3 = Inactivo
 - 4 = Menor de 10 años
- CAT_OCUP: CATEGORÍA OCUPACIONAL (Para ocupados y desocupados con ocupación anterior)
 - 1 = Patrón
 - 2 = Cuenta propia
 - 3 = Obrero o empleado,
 - 4 = Trabajador familiar sin remuneración
 - 9 = Ns./Nr.
- CAT_INAC: CATEGORÍA DE INACTIVIDAD
 - 1 = Jubilado/ Pensionado
 - 2 = Rentista
 - 3 = Estudiante
 - 4 = Ama de casa
 - 5 = Menor de 6 años
 - 6 = Discapacitado
 - 7 = Otros
- PP02H: En los últimos 12 meses ¿buscó trabajo en algún momento?
 - 1 = Si
 - 2 = No
- PP02I: En los últimos 12 meses ¿ha trabajado en algún momento?

- 1 = Si
 - 2 = No
- P47T: Monto de ingreso total individual.
- PONDII: Ponderador para ingreso total individual.
- V3_M: Monto del ingreso por indemnización por despido
- V5_M: Monto del ingreso por subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.
- ITF: Monto del ingreso total familiar
- IPCF: Monto del ingreso per capita familiar
- PONDIH: Ponderador del ingreso total familiar y del ingreso per cápita familiar, para hogares.

Cada registro tiene un número de identificación (CODUSU), que permite relacionar una vivienda con los hogares y personas que la componen a lo largo de los cuatro trimestres en que participa. En la base hogar todos los hogares que pertenecen a una misma vivienda poseen el mismo CODUSU. Para identificar los hogares se debe utilizar CODUSU y NRO_HOGAR. En la de personas todos los miembros del hogar tienen el mismo CODUSU y NRO_HOGAR pero se diferencian por el número de COMPONENTE.

Al querer estudiar la evolución económica de los individuos en la pandemia nos pareció lo mejor utilizar como variable central para el análisis el monto de ingreso total familiar (ITF). También se deberá contrastar con el monto de ingreso por subsidio o ayuda social (V5_M) que nos demostraría el peso de los planes o subsidios del estado. A raíz del plan de doble indemnización ofrecido por el gobierno nos parece interesante incluir al análisis la variable monto del ingreso por indemnización por despido (V3_M).

Se tiene en cuenta que se debe hacer una ponderación de los datos para su correcto análisis. Los campos PONDII, PONDIIO, PONDIH con corrección por no respuesta. Se utiliza PONDII para el tratamiento del ingreso total individual (p47t, decindr, adecindr, rdecindr, pdecindr, gdecindr, idecindr), PONDIIO para el ingreso de la ocupación principal (p21, pp06c, pp06d, pp08d1, pp08d4, pp08f1, pp08f2, pp08j1, pp08j2, pp08j3, decocur, adecocur, rdecocur, pdecocur, gdecocur, idecocur) y PONDIH para el ingreso total familiar (ITF, decifr, adecifr, rdecifr, pdecifr, gdecifr, idecifr), el ingreso per cápita familiar (IPCF, deccfr, adecifr, rdecifr, pdecifr, gdecifr, idecifr). El campo PONDERA, sin corrección, que se utiliza además para el resto de las variables.

Debido a que esta base es una muestra de la población Argentina se debe normalizar para que sea representativa a la misma. Para eso se utilizan los ponderadores PONDERA, PONDII y PONDIH. Estos se calculan para cada individuo y demuestran qué tan representativos son para el resto de la población, es decir, si una persona encuestada tiene PONDII 50 quiere decir que

el ingreso individual que esa persona representa a 50 individuos de toda la población. Identificamos también que existen ponderadores 0 esto se debe a que no todos los individuos responden la totalidad de la encuesta. En el caso de que los ponderadores sean 0 las variables que se ponderan a partir de ellas son -9, que es el default para decir que no se respondió la encuesta.

Para hacer las ponderaciones nos cruzamos con dos métodos: repeticiones y multiplicación. El método de ponderación mediante la repetición expande la muestra. Es decir, en el caso de que se esté normalizando la variable de monto de ingreso individual usando el ponderador PONDII, se repetirá PONDII veces el ingreso de esa persona. Este método no se masifican los datos, si no que se aumenta en su proporción. En cambio, el método de multiplicación masifica la variable de un individuo por todos los individuos que representa. Es decir, continuando con el ejemplo anterior, en el caso de que el ingreso individual de la persona sea de \$10 y represente a 10 personas, su ingreso ponderado se denotará como de $\$10 * 10 = 100$. Se debe tener más cuidado con el análisis que se lleva a la hora de hacer comparaciones porque al ser una variable asociada a la personas encuestada no se puede concluir que su ingreso es de \$100, sería un error muy grave de análisis.

Para entenderlo mejor mediante un ejemplo suponemos que queremos calcular la cantidad de individuos que tienen un ingreso mayor a \$50000, si utilizáramos el método de repetición nos devolverá la cantidad representativa a la población y sería beneficiosa para el análisis. Para utilizar el método de multiplicación primero se debería ponderar la variable condición porque si no no haría sentido para el análisis Pero el método de multiplicación no hace sentido al análisis porque la masificación del ingreso de esa persona a su PONDII no tiene relevancia con el ingreso de una persona. Para corregir este análisis se debe comparar en el ingreso ponderado por la condición ponderada y en el caso que se cumpla se sumariza el ponderador, que es la cantidad de personas que representa.

Se entiende que hay que ser cuidadosos con las conclusiones que damos y como nos referimos a los datos.

Para determinar la calidad de la base de datos se utilizó la función df_status() (FunModeling). Bajo este estudio se demuestra qué existen valores igual a cero en los casos qué se demuestren montos o ponderadores (cómo se demostró anteriormente es correcto qué así sea). Pero las variables CAT_OCUP, CAT_INAC, PP02H y PP02I presentan gran cantidad de ceros cuando no deberían. Estas son variables qué representan categorías según un número qué no debería ser cero. En estos casos el informe explicativo determina qué se utiliza cero en los casos a los cuales no les corresponde la secuencia analizada. Por eso, en el estudio de las variables decidimos no tener en cuenta aquellas qué no correspondan apartando los ceros. No se presentan valores nulos lo qué facilita mucho el análisis.

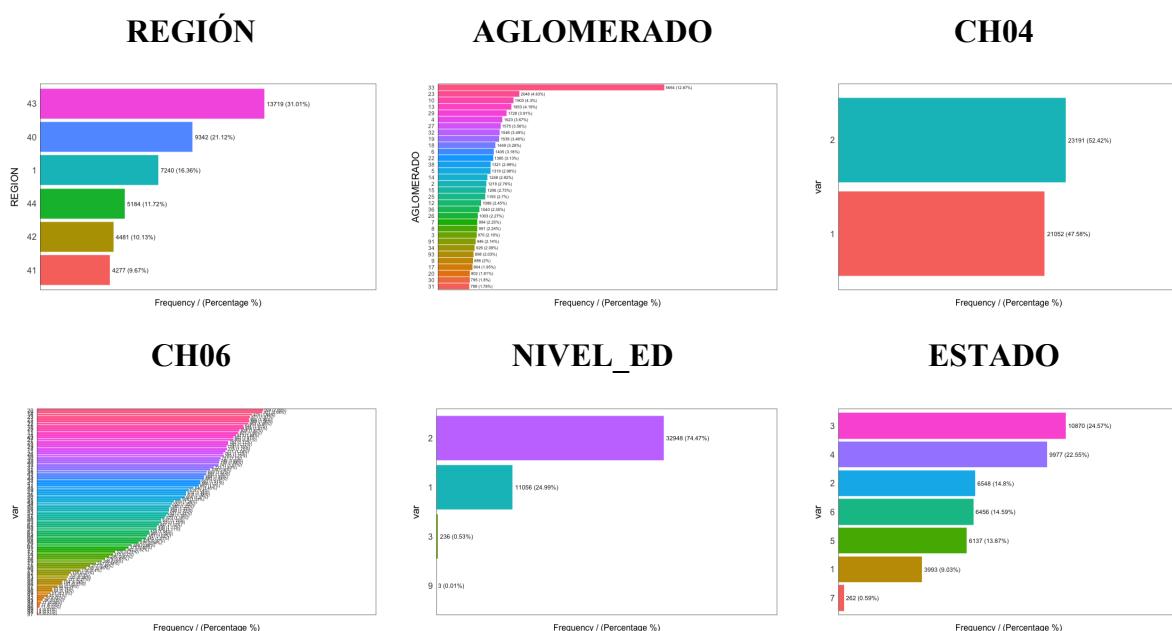
variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
	<int>	<dbl>	<int>	<dbl>	<int>	<dbl>	<chr>	<int>
CODUSU	0	0.00	0	0	0	0	factor	16663
NRO_HOGAR	0	0.00	0	0	0	0	integer	7
ANO4	0	0.00	0	0	0	0	integer	1
TRIMESTRE	0	0.00	0	0	0	0	integer	1
REGION	0	0.00	0	0	0	0	integer	6
AGLOMERADO	0	0.00	0	0	0	0	integer	32
PONDERA	0	0.00	0	0	0	0	integer	1134
CH04	0	0.00	0	0	0	0	integer	2
CH06	0	0.00	0	0	0	0	integer	90
CH10	0	0.00	0	0	0	0	integer	4
NIVEL_ED	0	0.00	0	0	0	0	integer	7
ESTADO	0	0.00	0	0	0	0	integer	3
CAT_OCUP	21503	48.60	0	0	0	0	integer	6
CAT_INAC	23150	52.32	0	0	0	0	integer	7
PP02H	23165	52.36	0	0	0	0	integer	3
PP02I	23150	52.32	0	0	0	0	integer	3
P47T	13090	29.59	0	0	0	0	integer	1213
PONDII	3732	8.44	0	0	0	0	integer	3138
V3_M	44181	99.86	0	0	0	0	integer	41
V5_M	41276	93.29	0	0	0	0	integer	296
ITF	8026	18.14	0	0	0	0	integer	1669
IPCF	8026	18.14	0	0	0	0	numeric	2663
PONDIH	7891	17.84	0	0	0	0	integer	2607

El dataset presenta una regla de default en la que se insertan los valores -9, 9, 99, 999 y 9999 para la variable "No Sabe / No Responde". Por eso, solo para el estudio de las variables decidimos convertir a nulos esos valores para que no afecten al análisis posterior. En caso de necesitarlo posteriormente volveremos a filtrar estas variables, pero no nos pareció conveniente ingresar valores nulos a toda la base sin sentido.

Variables categóricas

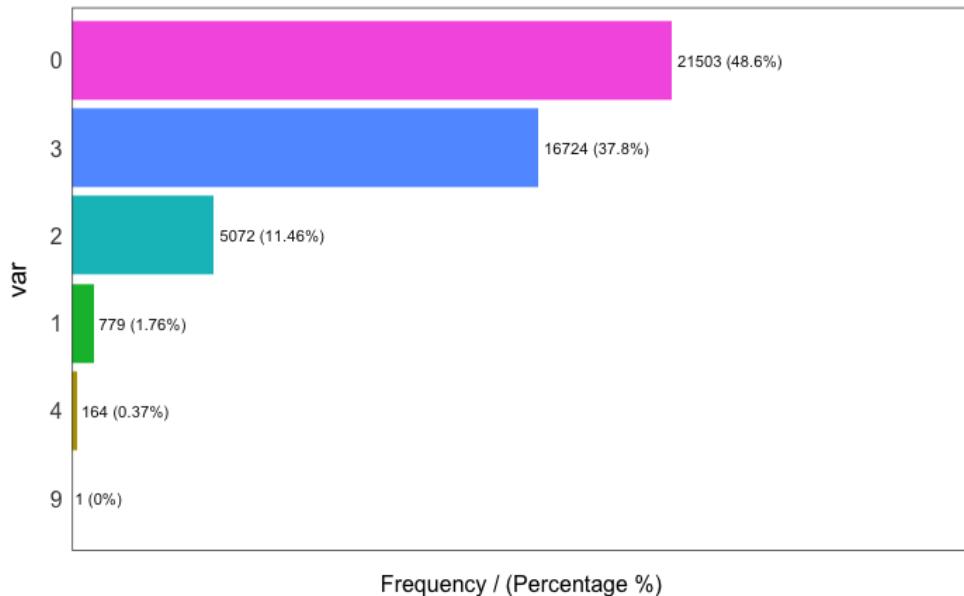
Para tener un panorama de cómo se distribuyen los datos en las variables categóricas usamos un gráfico de frecuencia del paquete funModelling. Las variables categóricas son: REGION, AGLOMERADO, CH04, CH06, CH10, NIVEL_ED, ESTADO, CAT_OCUP, CAT_INAC, PP02H, PP02I.

Estas fueron las que no presentaron anomalías (0 o NA):

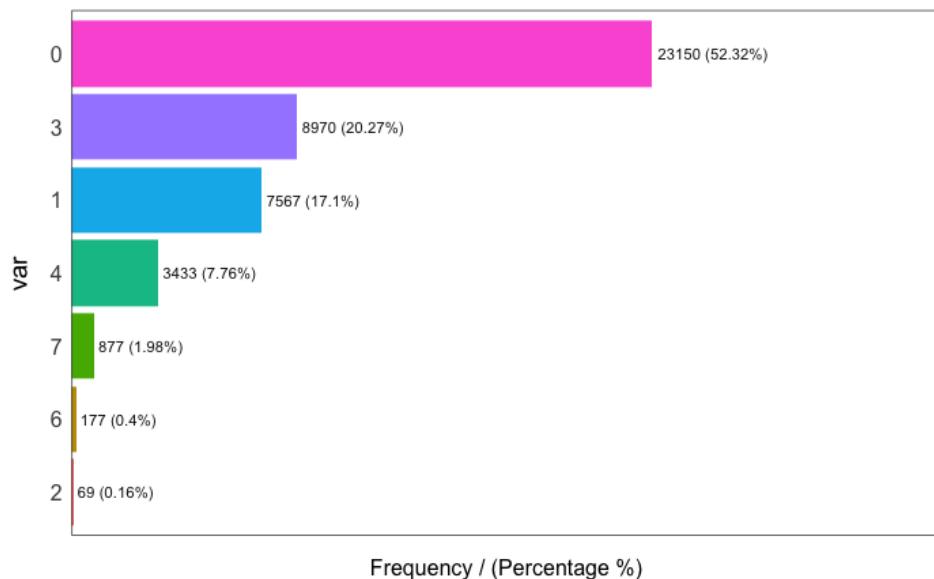


En el caso de CAT_OCUP, CAT_INAC, PP02H y PP02I si se presentaron 0, los cuales no representaban ninguna de las categorías especificadas en la nota metodológica, luego de investigar vimos que este 0 representa a las persona a las cuales no les corresponde contestar esa pregunta por alguna condición particular. En específico para cada variable significa:

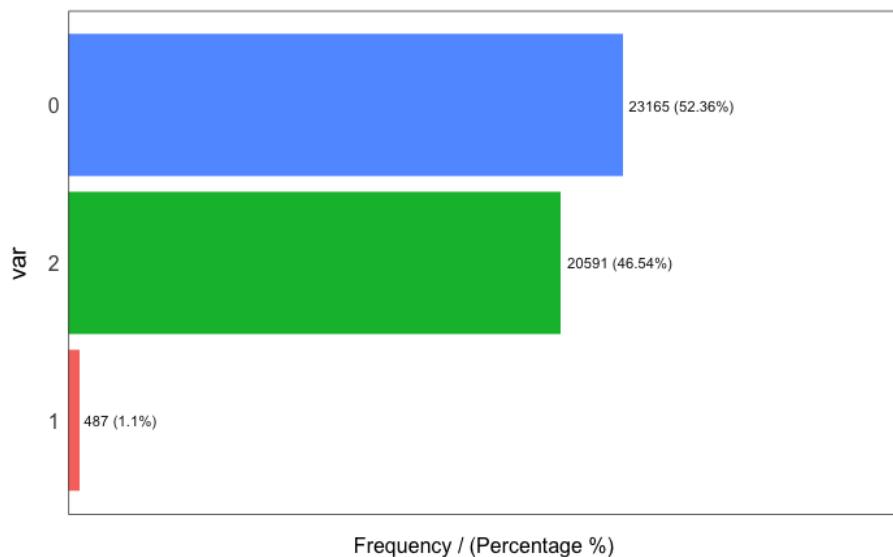
- CAT_OCUP: 0 significa que la persona encuestada está desocupada y nunca tuvo una ocupación anterior.



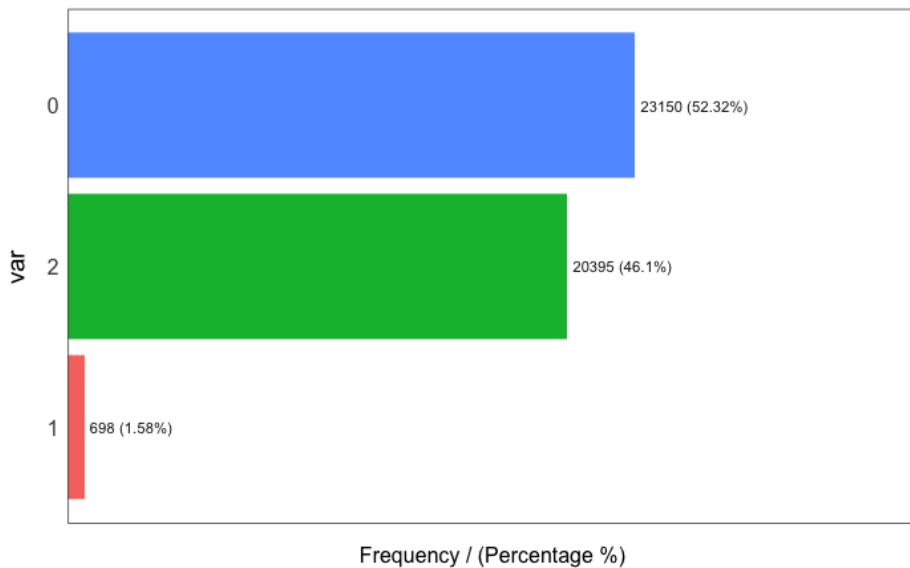
- CAT_INAC: 0 significa que la persona tiene ocupación



- PP02H: 0 significa que la persona tiene ocupación



- PP02I: 0 significa que la persona tiene ocupación hace más de 12 meses



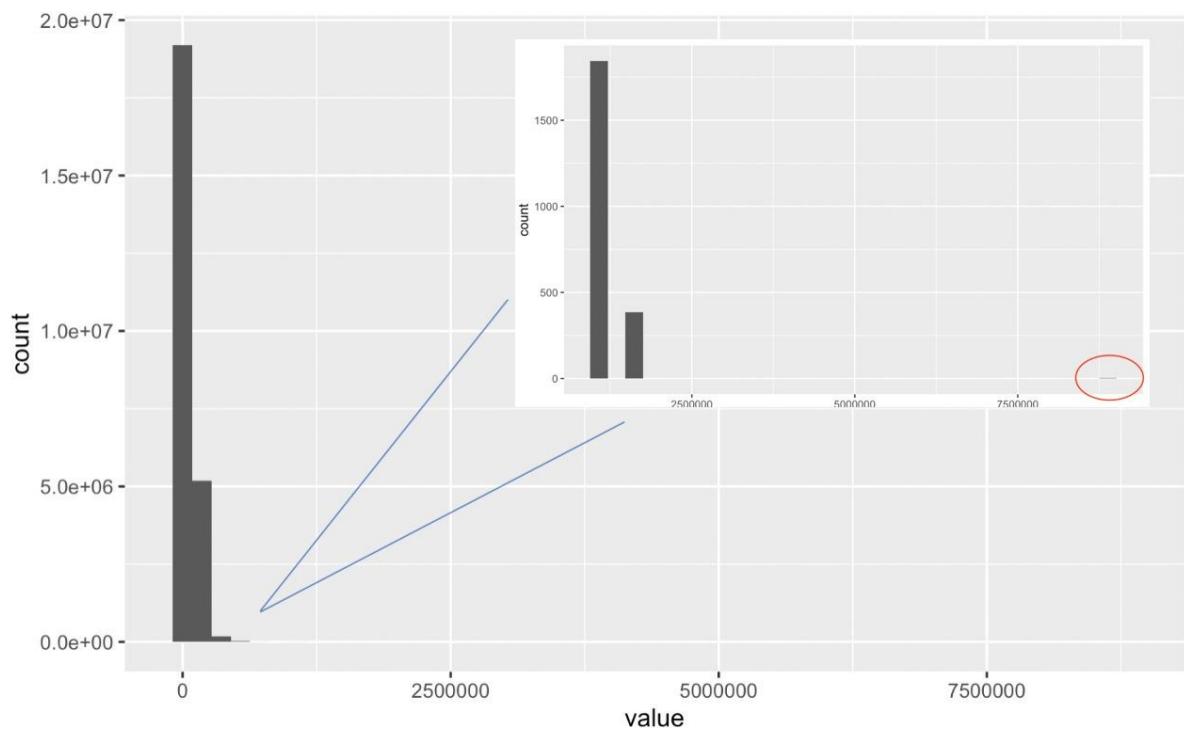
SUMMARY

BASE	Variable	Variable Explicativa	Mínimo	Primer Quartil	Mediana	Promedio	Tercer Quartil	Maximo
df_2020_T1	CODUSU	Código para distinguir VIVIENDAS						
	NRO_HOGAR	Código para distinguir HOGARES	1	1	1	1,048	1	51
	COMPONENTE	Nº de orden que se asigna a las personas que conforman cada hogar de la vivienda.						
	ANO4	Año de relevamiento	2020	2020	2020	2020	2020	2020
	TRIMESTRE		1	1	1	1	1	1
	REGION	Código de Region	1	40	42	35,32	43	44
	AGLOMERADO		2	10	22	23,23	32	93
	PONDERA		23	164	283	555	587	9628
	CH04	Sexo	1	1	2	1,524	2	2
	CH06	Edad	10	23	38	39,89	55	99
	CH10	¿Asiste o asistió a algún establecimiento educativo?	1	2	2	1,756	2	9
	NIVEL_ED	Nivel Educativo	1	3	4	3,636	5	7
	ESTADO	Condición de actividad	1	1	2	1,999	3	3
	CAT_OCUP	Categoría Ocupacional	0	0	1	1,396	3	9
	CAT_INAC	Categoría de Inactividad	0	0	0	1,256	3	7
	PP02H	En los últimos 12 meses ¿buscó trabajo en algún momento?	0	0	0	0,9418	2	2
	PP02I	En los últimos 12 meses ¿trabajó en algún momento?	0	0	0	0,9377	2	2
	p47T	Monto de ingreso total individual	-9	0	11500	18839	28000	8760000
	PONDII	Ponderador para ingreso total individual.	0	148	260	555,8	542	15258
	V3_M	Monto del ingreso por indemnización por despido	-9	0	0	182	0	1500000
	V5_M	Monto del ingreso por subsidio o ayuda social	-9	0	0	351,7	0	45000
	ITF	Monto del ingreso total familiar	0	18000	40000	50106	69000	8805000
	IPCF	Monto del ingreso per cápita familiar	0	4633	11320	16019	21250	4402500
	PONDIH	Ponderador del ingreso total familiar y del ingreso per capita familiar, para hogares.	0	126	233	555,8	474	17649

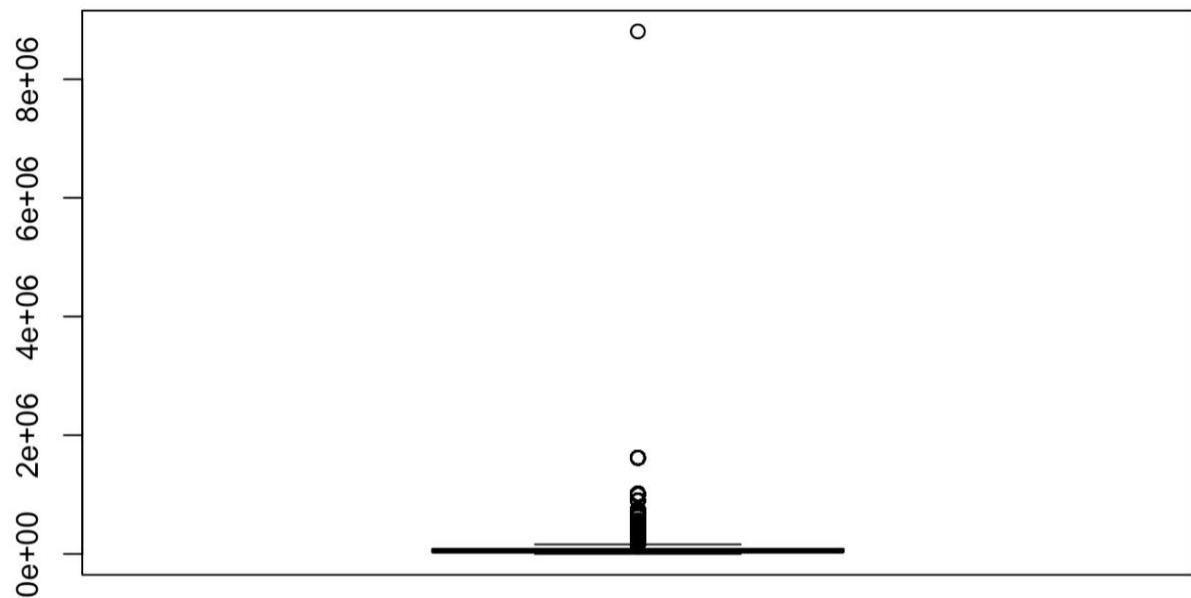
BASE	Variable	Variable Explicativa	Minimo	Primer Quartil	Mediana	Promedio	Tercer Quartil	Maximo
	CODUSU	Código para distinguir VIVIENDAS						
	NRO_HOGAR	Código para distinguir HOGARES	1	1	1	1,041	1	61
df_2020_T3	COMPONENTE	Nº de orden que se asigna a las personas que conforman cada hogar de la vivienda.						
	AN04	Año de relevamiento	2020	2020	2020	2020	2020	2020
	TRIMESTRE		3	3	3	3	3	3
	REGION	Codigo de Region	1	40	42	37,68	43	44
	AGLOMERADO		2	10	19	22,89	29	93
	PONDERA		21	140	240	668	526	21679
	CH04	Sexo	1	1	2	1,524	2	2
	CH06	Edad	10	23	38	40,12	55	102
	CH10	¿Asiste o asistió a algún establecimiento educativo?	1	1	2	1,755	2	3
	NIVEL_ED	Nivel Educativo	1	3	4	3,631	5	7
	ESTADO	Condición de actividad	1	1	3	2,064	3	3
	CAT_OCUP	Categoría Ocupacional	0	0	0	1,296	3	9
	CAT_INAC	Categoría de Inactividad	0	0	1	1,364	3	7
	PP02H	En los últimos 12 meses ¿buscó trabajo en algún momento?	0	0	1	0,9966	2	2
	PP02I	En los últimos 12 meses ¿trabajó en algún momento?	0	0	1	0,9875	2	2
	P47T	Monto de ingreso total individual	-9	0	12900	192116	28000	650000
	PONDII	Ponderador para ingreso total individual.	0	127	226	668,1	480	28920
	V3_M	Monto del ingreso por indemnización por despido	-9	0	0	56,98	0	240000
	V5_M	Monto del ingreso por subsidio o ayuda social	-9	0	0	1468	0	51000
	ITF	Monto del ingreso total familiar	0	21900	45000	53191	74900	691400
	IPCF	Monto del ingreso per cápita familiar.	0	7080	17000	336796	50000	23046667
	PONDIH	Ponderador del ingreso total familiar y del ingreso per capita familiar, para hogares.	0	102	213	668,1	441	21051

BASE	Variable	Variable Explicativa	Minimo	Primer Quartil	Mediana	Promedio	Tercer Quartil	Maximo
	CODUSU	Código para distinguir VIVIENDAS						
	NRO_HOGAR	Código para distinguir HOGARES	1	1	1,042	1	1	72
df_2021_T1	COMPONENTE	Nº de orden que se asigna a las personas que conforman cada hogar de la vivienda.						
	AN04	Año de relevamiento	2021	2021	2021	2021	2021	2021
	TRIMESTRE		1	1	1	1	1	1
	REGION	Codigo de Region	1	40	42	38,32	43	44
	AGLOMERADO		2	10	19	22,73	29	93
	PONDERA		18	155	252	610,9	473	16406
	CH04	Sexo	1	1	2	1,526	2	2
	CH06	Edad	10	23	38	40,52	56	101
	CH10	¿Asiste o asistió a algún establecimiento educativo?	1	2	2	1,759	2	9
	NIVEL_ED	Nivel Educativo	1	3	4	3,656	5	7
	ESTADO	Condición de actividad	1	1	2	2,017	3	3
	CAT_OCUP	Categoría Ocupacional	0	0	1	1,359	3	9
	CAT_INAC	Categoría de Inactividad	0	0	0	1,258	3	7
	PP02H	En los últimos 12 meses ¿buscó trabajo en algún momento?	0	0	0	0,9613	2	2
	PP02I	En los últimos 12 meses ¿trabajó en algún momento?	0	0	0	0,9589	2	2
	P47T	Monto de ingreso total individual	-9	0	14000	22823	33600	819000
	PONDII	Ponderador para ingreso total individual.	0	134	232	610,9	448	20452
	V3_M	Monto del ingreso por indemnización por despido	-9	0	0	86,9	0	600000
	V5_M	Monto del ingreso por subsidio o ayuda social	-9	0	0	541,7	0	250000
	ITF	Monto del ingreso total familiar	0	20000	50000	61666	85000	819000
	IPCF	Monto del ingreso per cápita familiar.	0	5800	14400	19451	25625	819000
	PONDIH	Ponderador del ingreso total familiar y del ingreso per capita familiar, para hogares.	0	110	212	610,9	425	22856

Para estudiar valores extremos separamos las variables categóricas de las cuantitativas. La mayoría de las variables son categóricas, por lo que no se puede hacer análisis de outliers sobre ellas. Las variables existentes cuantitativas son montos de ingreso (P47T, V3_M, V5_M, ITF y IPCF). Estas variables siguen una distribución lognormal por lo que hay muchas variables de valores bajos y muy pocas de valores muy altos. Como podemos ver en el gráfico:



Por esto, si estudiamos los valores extremos encontraríamos muchos de ellos ya que los valores máximos tienden a ser muy altos y alejados del tercer cuartil.



Es por eso que decidimos no normalizar los valores extremos ya que no consideramos que realmente lo sean y queremos estudiar en la realidad los ingresos.

Correlación entre variables

Para el análisis de correlaciones entre variables se separaron las variables numéricas. Todas las variables menos el CODUSU son numéricas, pero no todas ellas son cuantitativas y estaría mal

por ejemplo inferir que a mayor sea la región (en número) mayor es el ponderador. Por eso, solo dejamos las variables cuantitativas. No se ponderan las variables porque tendrían una correlación falsa con las variables ponderadas.

	P47T	V3_M	V5_M	ITF	IPCF	PONDERA	PONDII	PONDIH
P47T	1.0000000000	0.212410525	-0.017264804	0.64833324	0.704725877	0.0009003412	0.054636951	0.047828597
V3_M	0.212410525	1.0000000000	-0.003514316	0.14084483	0.152474553	-0.0040814300	-0.002530038	-0.002214387
V5_M	-0.0172648043	-0.003514316	1.0000000000	-0.03867107	-0.051366429	-0.0173537138	-0.003668998	-0.005518908
df_2020_T1	ITF	0.6483332440	0.140844829	-0.038671066	1.00000000	0.909056181	-0.044592468	0.031585216
IPCF	0.7047258769	0.152474553	-0.051366429	0.90905618	1.00000000	-0.0083639091	0.051195680	0.079950032
PONDERA	0.0009003412	-0.004081430	-0.017353714	-0.04459247	-0.008363909	1.0000000000	0.838095808	0.681577980
PONDII	0.0546369507	-0.002530038	-0.003668998	0.03158522	0.051195680	0.8380958084	1.000000000	0.807418772
PONDIH	0.0478285973	-0.002214387	-0.005518908	0.10528126	0.079950032	0.6815779802	0.807418772	1.000000000
	P47T	V3_M	V5_M	ITF	IPCF	PONDERA	PONDII	PONDIH
P47T	1.0000000000	0.0897699269	0.0003527272	0.50363330	0.1698240537	-0.013436430	0.0714594246	0.0667894757
V3_M	0.0897699269	1.0000000000	0.0002752139	0.06380772	0.0236809329	-0.001545581	-0.0003463411	0.0010609303
V5_M	0.0003527272	0.0002752139	1.0000000000	-0.03780980	-0.0006455069	-0.008669394	0.0075801004	-0.0004836051
df_2020_T3	ITF	0.5036333039	0.0638077171	-0.0378098045	1.00000000	0.3622899885	-0.075297087	0.0103323341
IPCF	0.1698240537	0.0236809329	-0.0006455069	0.36228999	1.0000000000	-0.043215456	-0.0155156603	0.0183013378
PONDERA	-0.0134364305	-0.0015455806	-0.0086693942	-0.07529709	-0.0432154560	1.0000000000	0.8852009812	0.7441696647
PONDII	0.0714594246	-0.0003463411	0.0075801004	0.01033233	-0.0155156603	0.885200981	1.0000000000	0.8369305528
PONDIH	0.0667894757	0.010609303	-0.0004836051	0.11011258	0.0183013378	0.744169665	0.8369305528	1.0000000000
	P47T	V3_M	V5_M	ITF	IPCF	PONDERA	PONDII	PONDIH
P47T	1.0000000000	0.138617066	0.014784610	0.531138836	0.67256032	-0.006266979	0.084109627	0.078657030
V3_M	0.138617066	1.0000000000	0.002057940	0.042263322	0.03731142	0.005002784	0.007961630	0.001292373
V5_M	0.014784610	0.002057940	1.0000000000	-0.009660039	-0.02777683	-0.010658224	-0.004779385	-0.006904138
df_2021_T1	ITF	0.531138836	0.042263322	-0.009660039	1.0000000000	0.76395461	-0.027877873	0.059211783
IPCF	0.672560320	0.037311417	-0.027776826	0.763954614	1.0000000000	-0.022645878	0.058855026	0.123467602
PONDERA	-0.006266979	0.005002784	-0.010658224	-0.027877873	-0.02264588	1.0000000000	0.878794351	0.751639848
PONDII	0.084109627	0.007961630	-0.004779385	0.059211783	0.05885503	0.878794351	1.000000000	0.848127500
PONDIH	0.078657030	0.001292373	-0.006904138	0.164233645	0.12346760	0.751639848	0.848127500	1.000000000

En su mayoría, las variables no presentan altos niveles de correlación. Sí se encuentra relación entre los ponderadores:

- Entre PONDERA y PONDII se encuentra correlación positiva en los tres trimestres.
- Entre PONDII y PONDIH se encuentra correlación positiva en los tres trimestres.

En el caso de la correlación entre el Ingreso Total Familiar (ITF) y el Monto de Ingreso Per Capita Familiar (IPCF) en el primer trimestre del 2020 tienen alta correlación y luego en los próximos esa correlación disminuye. Nos llamó mucho la atención porque qué están correlacionados significa qué a mayor sea el Ingreso Total mayor sería el Ingreso per Cápita, qué tiene lógica.

Cómo las variables no se encuentran correlacionadas, no correspondería hacer análisis de correspondencias simples ni múltiples (PCA y MCA).

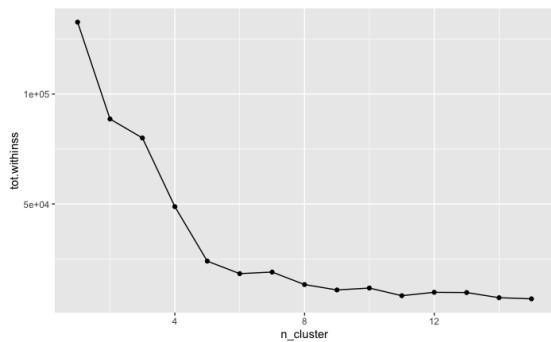
Clusters

Para hacer clusters, al tener tantas variables, las bases son muy pesadas y se acaba la memoria rápidamente. Más en el caso de hacer la ponderación de las mismas. Por eso decidimos, teniendo en cuenta que va a ser un análisis de la muestra y no del total de la población, hacer el análisis sin hacer ponderaciones, identificar que variables repercuten en mayor medida en las agrupaciones e identificar si estos clusters se repiten en las variables seleccionadas y ponderadas.

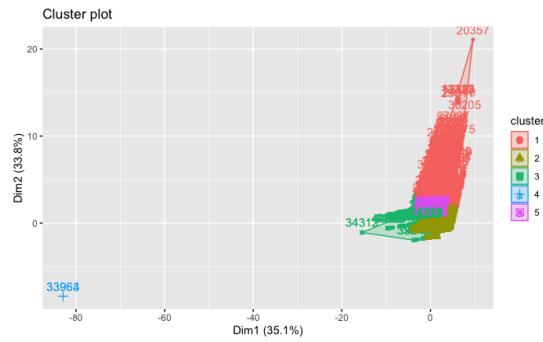
Al intentar hacer clusters a partir del dataset, encontramos varias dificultades, por lo que decidimos filtrarlo y quedarnos solo con las variables Región, ITF y V5_M.

2020 T1

Cantidad de clusters: 5

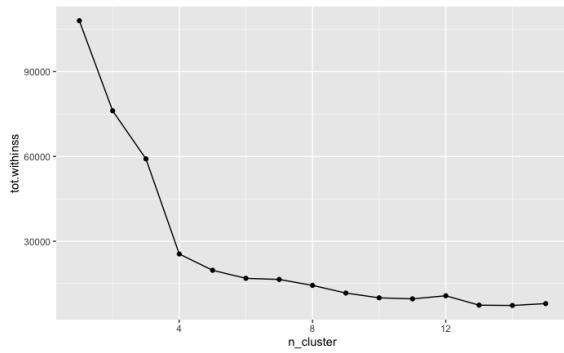


Clusters graficados

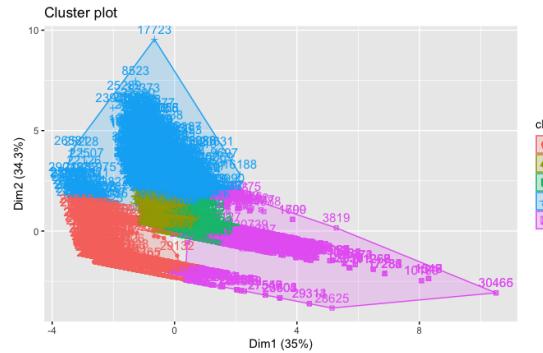


2020 T3

Cantidad de clusters: 5

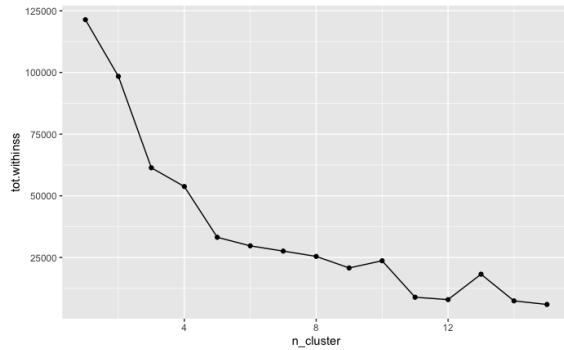


Clusters graficados

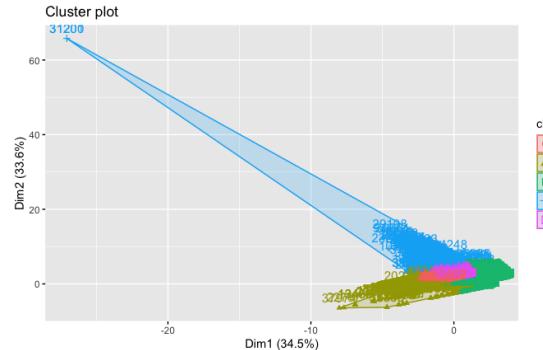


2021 T1

Número de clusters: 5



Clusters graficados



Analizamos puntualmente el tercer trimestre (Octubre-Noviembre-Diciembre) de 2020, plena pandemia, agregando más variables al análisis y generando nuevos clusters.

2020 T3

Variables: REGION, CH04, CH06, NIVEL_ED, ESTADO, V3_M, V5_M, ITF

Cantidad de clusters: 4

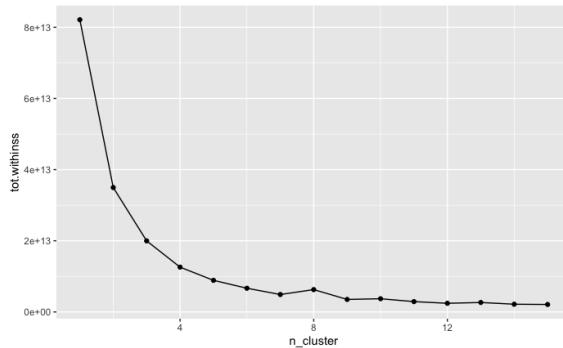
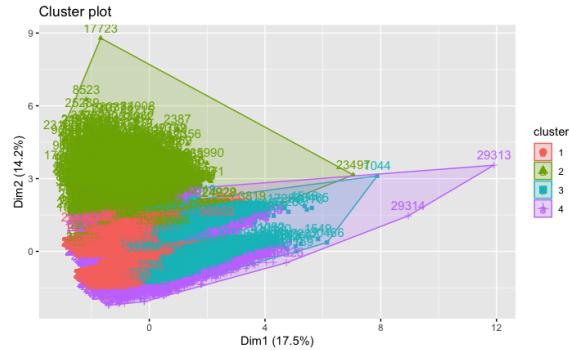
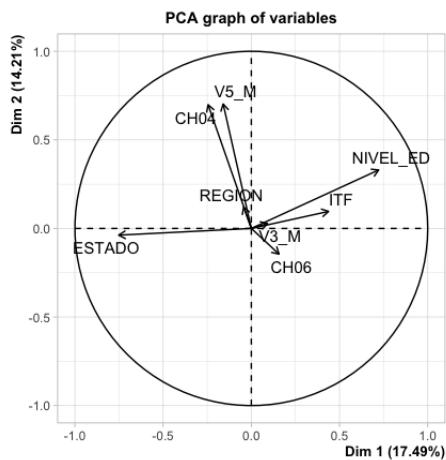


Gráfico de clusters:

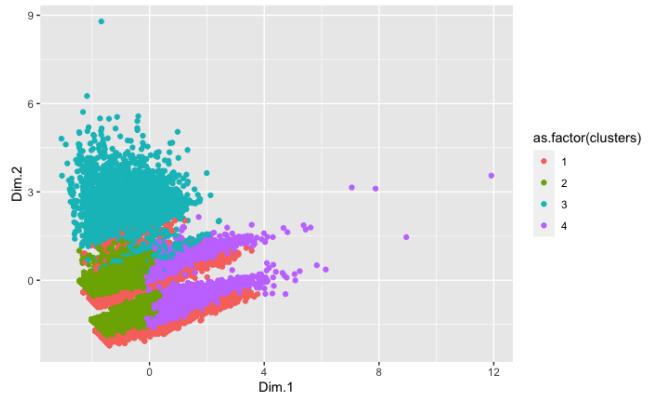


Al observar que los clusters formados se superponen y el gráfico presenta mucho “ruido”, usamos la técnica de combinar clusters con PCA (Principal Component Analysis) para ver si esto reduce el *overlap* y nos permite sacar conclusiones.

Resultados del PCA



PCA + Clusters

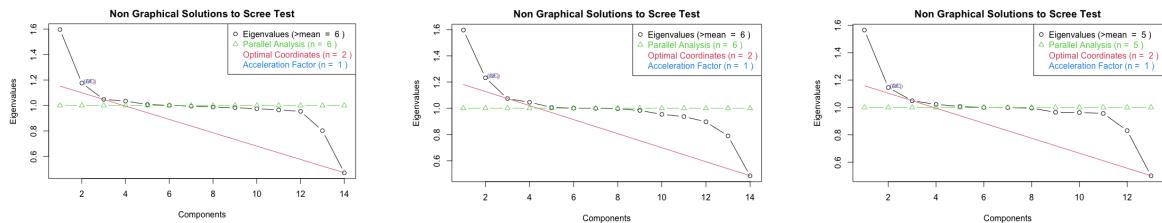


Por más qué se combinó clusters con PCA, los resultados no mejoraron. Los clusters no resultaron significativos ya qué sigue existiendo overlap y mucho “ruido” en el gráfico.

Análisis Factorial

Otra cuestión a explorar es el impacto de los distintos ingresos de las familias en el ingreso total en los tres períodos, previo, durante y después de la pandemia. Para esto decidimos hacer análisis factorial.

El análisis factorial busca encontrar uno o varios factores latentes (invisibles) que están afectando los datos. Se concluye qué existe un único factor, según el test realizado, por lo tanto los análisis se hicieron con nfactors = 1.



Agregamos las siguientes variables para el análisis factorial:

- P21 = Monto de ingreso de la ocupación principal.
- TOT_P12 = Monto de ingreso de otras ocupaciones
- V2_M = Monto del ingreso por jubilacion o pension
- V4_M = Monto del ingreso por seguro de desempleo
- V8_M = Monto del ingreso por alquiler de su propiedad.
- V9_M = Monto del ingreso por ganancias de algún negocio en el que no trabaja
- V10_M = Monto del ingreso por intereses o rentas por plazos fijos/inversiones
- V11_M = Monto del ingreso por BECA DE ESTUDIO.
- V12_M = Monto del ingreso por cuotas de alimentos o ayuda en dinero de personas que no viven en el hogar.
- V18_M = Monto del ingreso por otros ingresos en efectivo
- V19_AM = Monto del ingreso por trabajo de menores de 10 años. (* no incluida en 2021 T1)
- V21_M = Monto del ingreso por aguinaldo.

Para investigar si el dataset era adecuado para realizar análisis factorial, usamos el criterio de KMO. Según este criterio el KMO debería ser mayor a 0.6 y los resultados en cada dataset fueron:

2020 T1

0.53

2020 T3

0.52

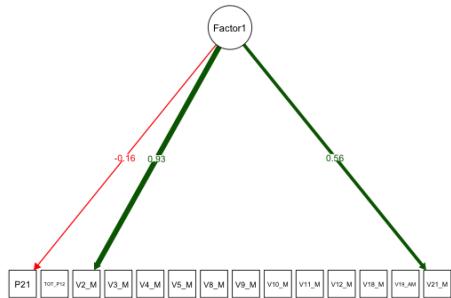
2021 T1

0.52

Los resultados son cercanos a lo deseados pero no óptimos, de todas maneras, continuamos con el análisis.

2020 T1

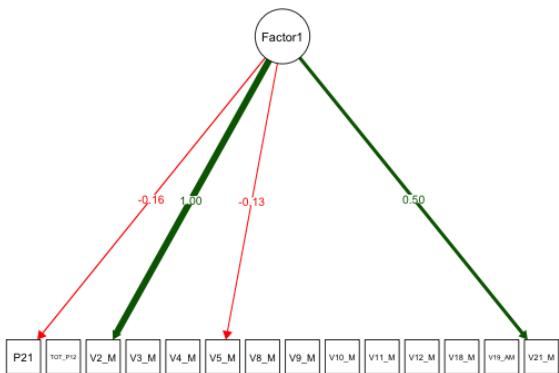
Análisis Factorial



- Positivamente V2_M: jubilación o pensión
- Positivamente en menor medida V21_M: aguinaldo
- Negativamente P21: monto de la ocupación principal

2020 T3

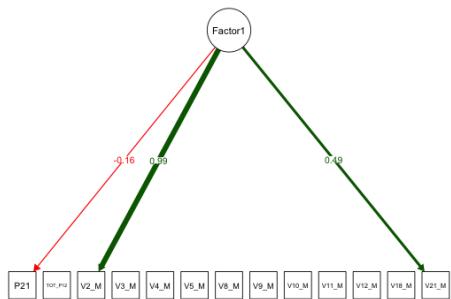
Análisis Factorial



- Positivamente V2_M: jubilación o pensión
- Positivamente en menor medida V21_M: aguinaldo
- Negativamente P21: monto de la ocupación principal
- Negativamente V5_M: planes sociales o ayudas económicas

2021 T1

Análisis Factorial



- Positivamente V2_M: jubilacion o pension
- Positivamente en menor medida V21_M: aguinaldo
- Negativamente P21: monto de la ocupación principal

Durante la pandemia hubo cambios en el factor, donde la variable V5_M (ayuda o planes sociales) impactó negativamente al Factor1.

MCA

Se quiere estudiar si las características qué describen a las personas qué toman planes son las mismas en períodos previos y durante la pandemia. Se estudiará sobre los datos de la muestra,

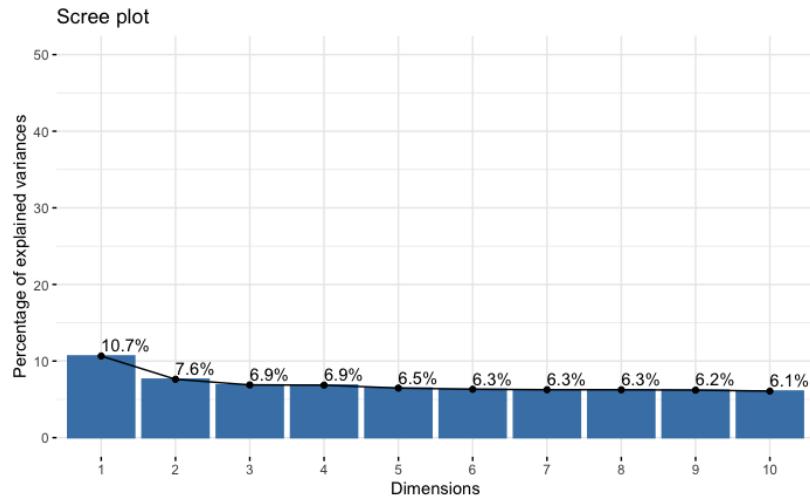
sin ponderación por lo qué no se puede inferir qué las conclusiones suceden en la realidad de la población.

Para las variables categóricas del tipo:

- V5 = ¿En los últimos tres meses, las personas de este hogar han vivido de subsidio o ayuda social (en dinero) del gobierno, iglesias, etc.?
 - 1 = Sí
 - 2 = No
- REGIÓN = Código de Región
 - 01 = Gran Buenos Aires
 - 40 = Noroeste
 - 41 = Nordeste
 - 42 = Cuyo
 - 43 = Pampeana
 - 44 = Patagónica
- II7 = Régimen de tenencia
 - 01 = Propietario de la vivienda y el terreno
 - 02 = Propietario de la vivienda solamente
 - 03 = Inquilino/arrendatario de la vivienda
 - 04 = Ocupante por pago de impuestos/expensas
 - 05 = Ocupante en relación de dependencia
 - 06 = Ocupante gratuito (con permiso)
 - 07 = Ocupante de hecho (sin permiso)
 - 08 = Está en sucesión?
- V1 = ...de lo que ganan en el trabajo?
 - 1 = Si
 - 2 = No
- V2 = ...de alguna jubilación o pensión?
 - 1 = Si
 - 2 = No

Se hizo análisis MCA, del cual obtuvimos resultados desalentadores ya que entre el primer y el segundo componente solo se explica el 18,3% de la varianza, lo cual significa que este no es un método apropiado para analizar este dataset.

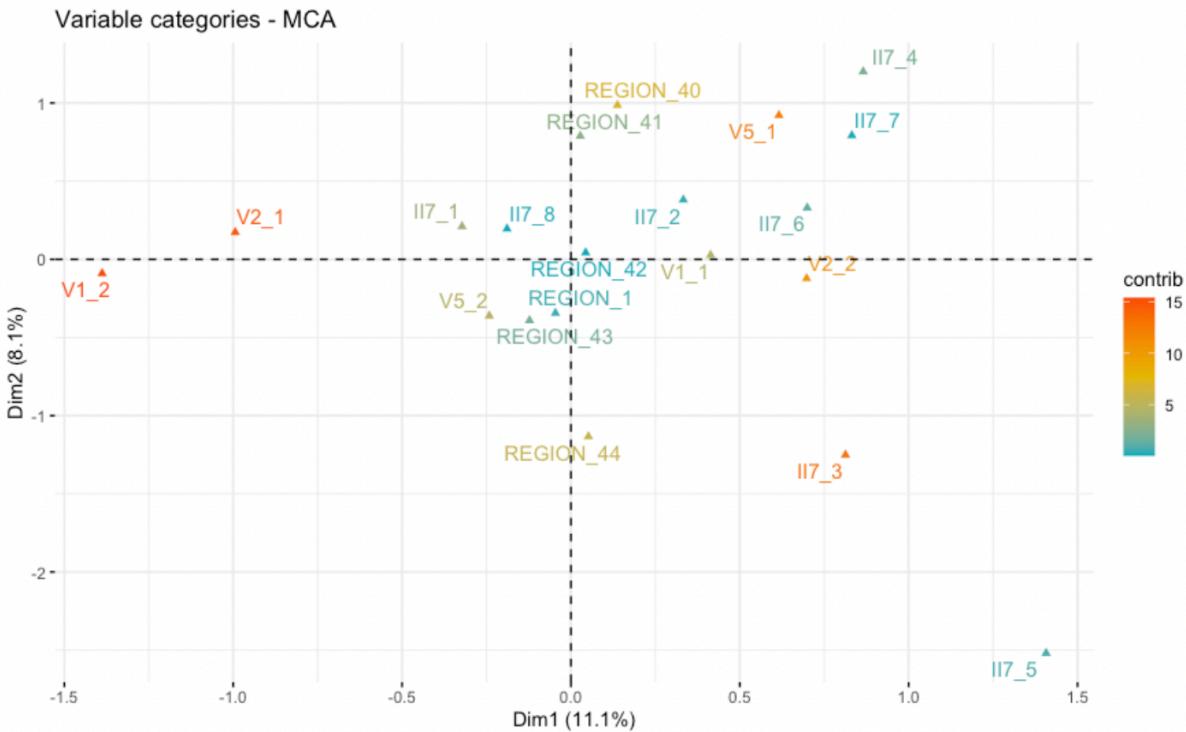
% de varianza explicada



Análisis general en pandemia

De todas maneras se puede identificar 4 grupos:

- V1_2 y V2_1 con Dim2 neutral y Dim1 < 0: No viven de lo que ganan en el trabajo y perciben una jubilación o pensión.
- Esquina superior derecha (II7_7, V5_1, II7_2, II7_4, II7_6, REGION_41, REGION_40) con Dim1 > 0 y Dim2 > 0: Las personas de este grupo son ocupantes de vivienda de hecho (sin permiso), propietario de la vivienda solamente, ocupante por pago de impuestos/expensas o ocupante gratuito (con permiso). Son personas qué en los últimos tres meses han vivido de subsidio social y de la región del Noroeste o Noreste.
- Esquina inferior derecha (II7_0, II7_3, II7_5) con Dim 2 > 0 y Dim 1 < 0: Son inquilinos u ocupante en relación de dependencia, deben estar al pendiente de un ingreso para pagar su vivienda.
- Neutro Dim 1 y Dim 2 cercanas a 0: (II7_1, Region_1, Region_42, Region_43, Region_44, V5_2, II7_8, II7_2, V1_1): En este grupo están personas de la región centro y pampeana, que vive de su trabajo y no recibe planes sociales



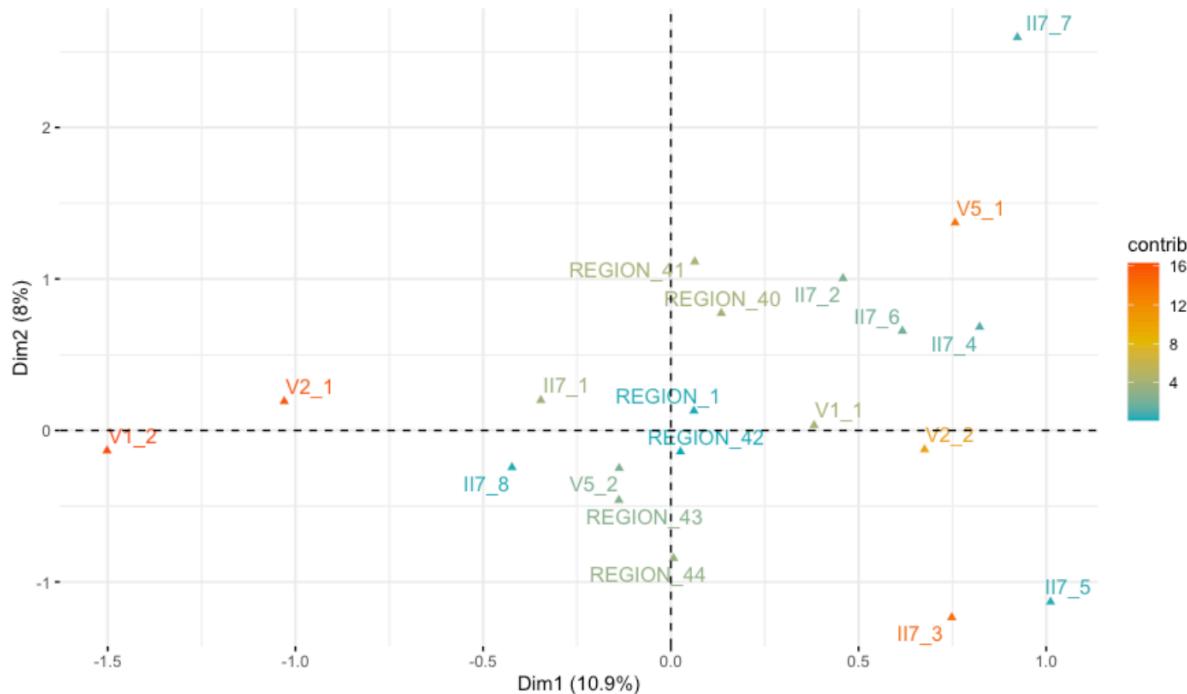
Podemos decir que las personas con alta Dim 1, son el grupo más afectado económicoamente por la pandemia, a quienes se apuntaron muchos planes sociales como el IFE . Esto incluye a personas que trabajan en relación de dependencia, que alquilan y no disponen de vivienda propia, además cobran planes sociales o viven en el norte de la Argentina (región con más pobreza).

Por otro lado, una baja Dim 2 muestra que básicamente los que están jubilados.

En el centro se encuentran personas estables económicoamente.

Análisis general pre-pandemia

Variable categories - MCA



Análisis MCA

1. **Nuevos sectores necesitan ayudas sociales:** Aceramiento general de las variables a V5_1 (recibe planes)
2. **Efecto del congelamiento de alquileres, prohibición de desalojo y suspensión de corte de servicios :** II7_3 (inquilino o arrendatario) estable debido a que no necesitan planes porque sus necesidades fueron enmendadas con las medidas anterior
3. **Impacto nacional:** Los indicadores de región se juntaron mostrando que la pandemia afectó a todo el país
4. **Ocupación de terrenos*:** Aceramiento de II7_7 (ocupación sin permiso) a V5_1 causado, no solo por el aumento de personas que ocuparon terrenos, si no también por acondicionamiento por parte del gobierno de los terrenos (agua, luz, cloaca) para que continuaran allí. *
5. **Por región:** Se segmentan las regiones en las que adquieren plan (se acercan a V5_1), Noroeste y Nordeste, y las qué no (se acercan a V5_2), Cuyo, Pampeana, Gran Buenos Aires y Patagónica.
6. **Prohibición de despido:** Alejamiento de II7_5 a V5_1, II7_5 representa a las personas que habitan una vivienda facilitada gratuitamente por la empresa donde trabaja, por lo tanto, no tienen necesidad de ayuda económica

Análisis de monto de planes sociales (Test de Hipótesis)

Se quiere conocer la variación del monto de ingreso por subsidio o planes sociales en los tres trimestres predeterminados. Para eso, ponderamos estas variables según su respectivo ponderador. En el caso de V5_M, al ser una proporción del Monto total de ingreso, se pondera usando la repetición de la variable según PONDIH.

Como los vectores son de distinto tamaño no se puede realizar la comparación entre las diferencias de medias de la población de datos (obtenida ponderando la base), por ende, se extrae una muestra aleatoria de igual tamaño de las tres bases y se hará un estudio para saber si hay una diferencia de monto de ingreso por subsidios estadísticamente significativos mediante una prueba de hipótesis.

Variables ponderadas:

a = V5_M del primer trimestre del 2020

b= V5_M del tercer trimestre del 2020

c = V5_M del primer trimestre del 2021

Variables:

X_1 = Monto de ingreso por subsidio (V5_M) en el primer trimestre del 2020.

X_2 = Monto de ingreso por subsidio (V5_M) en el tercer trimestre del 2020.

Hipótesis:

$H_0: \mu(a) = \mu(b)$

$H_1: \mu(a) \neq \mu(b)$

Condición de Rechazo: $p\text{-value} \leq 0.05$

Regla de decisión: Se toma una muestra y se calcula el p-value. Si ese valor es menor o igual a 0.05, se rechaza la hipótesis nula, es decir existen diferencias significativas entre las medias. En cambio, si es mayor, no se rechaza la hipótesis, lo que significa que no hay evidencia suficiente para decir que hay diferencias entre las medias.

Como el p-value < 0.0003, se rechaza H_0 , se ve una diferencia significativa entre los montos de ingreso del primer y tercer trimestre.

```
t = -3.6472, df = 248.47, p-value = 0.0003229
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-1553.9283 -464.1417
sample estimates:
mean of x mean of y
266.230 1275.265
```

Lo mismo se estudia para los períodos de plena pandemia (3er trimestre 2020) y el trimestre de relajo posterior a la pandemia (1er trimestre 2021).

Variables:

X_2 = Monto de ingreso por subsidio (V5_M) en el tercer trimestre del 2020.

X_3 = Monto de ingreso por subsidio (V5_M) en el primer trimestre del 2021.

Hipótesis:

$$H_0: \mu(b) = \mu(c)$$

$$H_1: \mu(b) \neq \mu(c)$$

Condición de Rechazo: $p\text{-value} \leq 0.05$

Regla de decisión: Se toma una muestra y se calcula el p-value. Si ese valor es menor o igual a 0.05, se rechaza la hipótesis nula, es decir existen diferencias significativas entre las medias. En cambio, si es mayor, no se rechaza la hipótesis, lo que significa que no hay evidencia suficiente para decir que hay diferencias entre las medias.

Como el p-value < 0.0016 , se rechaza H_0 , se ve una diferencia significativa entre los montos de ingreso del tercer trimestre del 2020 y el primer trimestre del 2021.

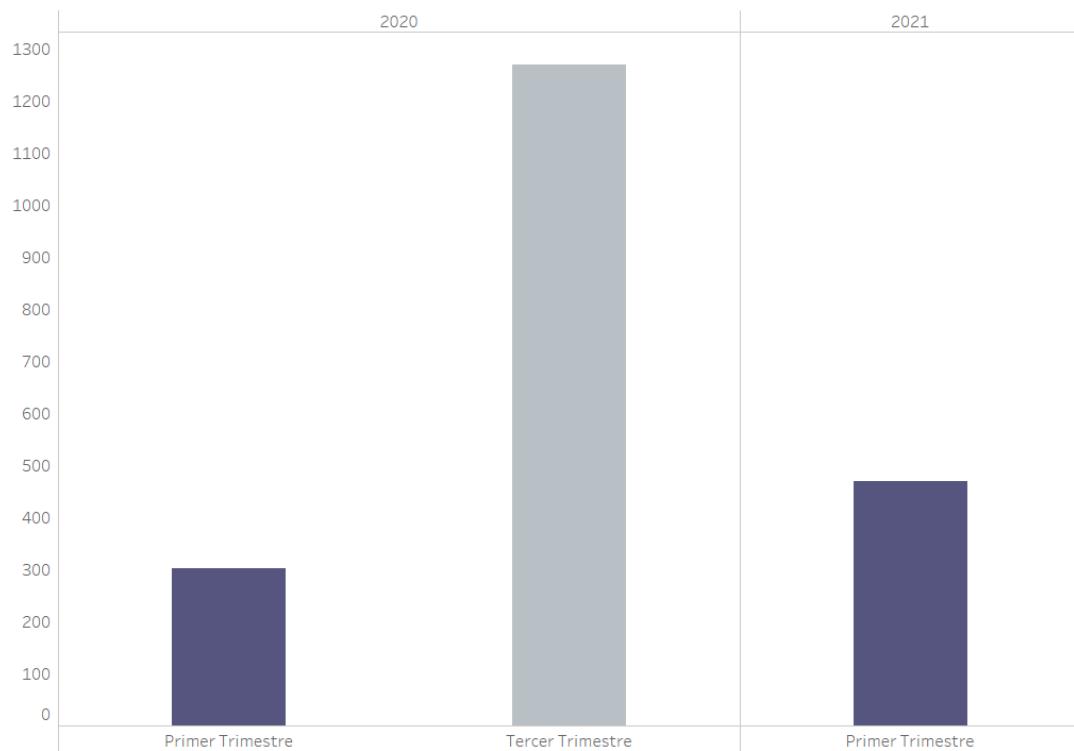
```
Welch Two Sample t-test

data: b and c
t = 3.1754, df = 307.15, p-value = 0.001648
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 358.3299 1526.0101
sample estimates:
mean of x mean of y
1275.265   333.095
```

A partir de este análisis se puede concluir qué hubo una clara diferencia a nivel poblacional del monto de ingreso por subsidio entre los distintos períodos.

Variación del Monto de Ingreso por Subsidios

Se desea interpretar la variación qué hubo en el monto de ingreso por subsidio total ponderado para ver si acompaña el análisis anterior. Se debería encontrar crecimiento entre el primer trimestre y el tercer trimestre del 2020, ya que el test de hipótesis demuestra diferencia significativa entre ellas y la media del último periodo es mucho mayor a la del primero. Además, esto acompaña la lógica de entendimiento de la realidad, donde se notó un fuerte impacto en la necesidad de planes sociales a partir de la pandemia. Es el caso inverso la variación entre el tercer trimestre del 2020 y el primero del 2021, debería disminuir el monto de ingreso por planes de uno al otro, ya que, el test de hipótesis demuestra diferencia significativa entre ellas y la media del último periodo es mucho menor a la del primero.



Análisis de proporción de población que recibe planes sociales

Para averiguar si la proporción de población que percibe planes sociales o ayudas económicas aumentó en durante la pandemia, se hizo un test de proporciones.

Variables:

p_{2020} : proporción de personas que reciben planes sociales en el 1º trimestre de 2020 ($V5=1$)

p_{2021} : proporción de personas que reciben planes sociales en el 1º trimestre de 2021 ($V5=1$)

Hipótesis:

H_0 : Durante la pandemia menos o igual cantidad de población recibió planes sociales ($p_{2020} \leq p_{2021}$)

H_1 : Durante la pandemia más gente recibió planes sociales ($p_{2020} > p_{2021}$)

Condición de Rechazo: $p-value \leq 0.05$

Regla de decisión: Se toma una muestra y se calcula el p-value. Si ese valor es menor o igual a 0.05, se rechaza la hipótesis nula, es decir durante la pandemia más gente recibe planes sociales. En cambio, si es mayor, no se rechaza la hipótesis, lo que significa que no hay evidencia suficiente para decir que durante la pandemia más gente recibe planes sociales.

Como el p-value es 0.006109, se rechaza la hipótesis nula, es decir que durante la pandemia más gente recibió ayuda económica.