

**82.04 - Analitica Descriptiva**

**Trabajo Práctico Final**

Segundo cuatrimestre de 2021

Collado, Camila (61487)

Noguera, Abril (61541)

Pettinato, Camila (61050)

[SCRIPT](https://drive.google.com/file/d/1ajquW8W1VZI-UeSgxQ3dNtSR0_E7B6sb/view?usp=sharing)

### 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 relajo 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 cómo 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 demostrara 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.

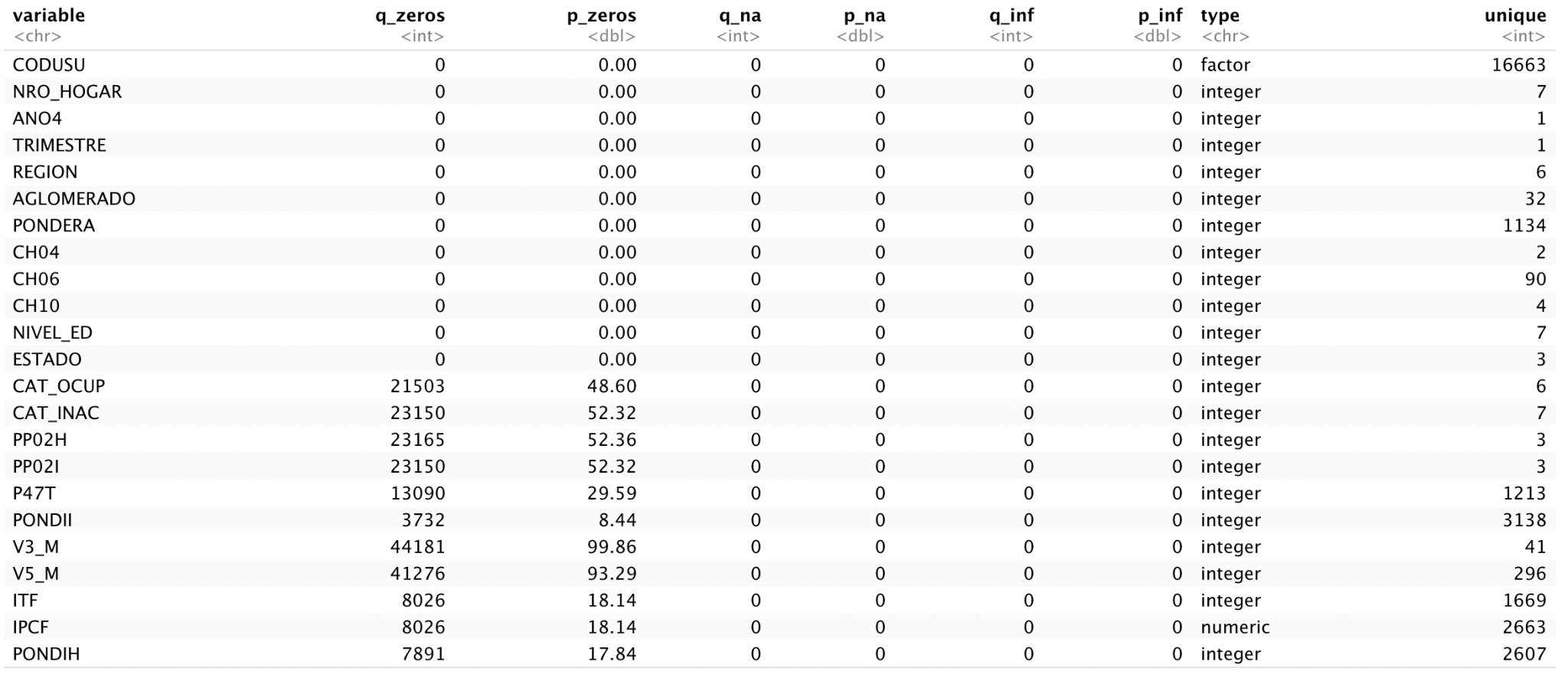
Debido a que esta base es una muestra de la población Argentina se debe normalizar para que sea representativa a la misma. Para eso se utilizan los ponderadores PONDERA, PONDII y PONDIH. Estos se calculan para cada individuo y demuestran qué tan representativos son para el resto de la población, es decir,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



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**

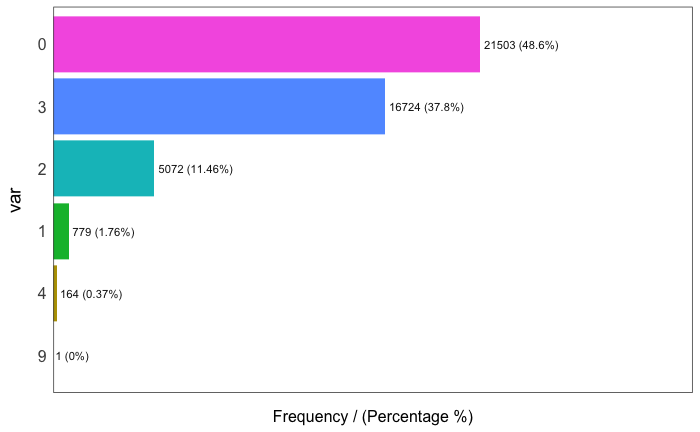
Para tener un panorama de cómo se distribuyen los datos en las variables categóricas usamos un gráfico de frecuencia del paquete funModelling. Las variables categóricas son: REGION, AGLOMERADO, CH04, CH06, CH10, NIVEL\_ED, ESTADO, CAT\_OCUP, CAT\_INAC, PP02H, PP02I.

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

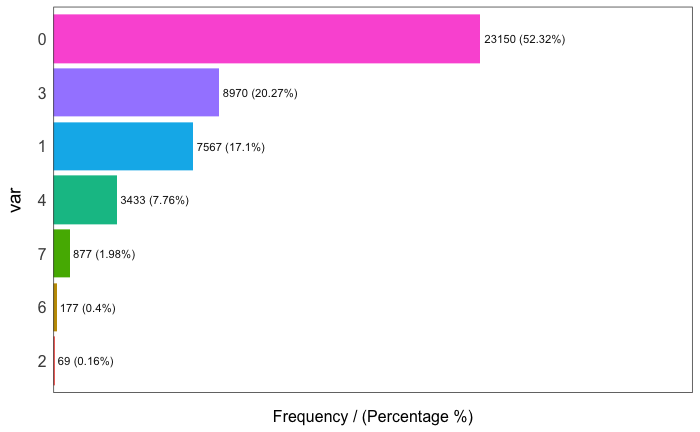
| **REGIÓN** | **AGLOMERADO** | **CH04** |
| --- | --- | --- |
|  |  |  |
| **CH06** | **NIVEL\_ED** | **ESTADO** |
|  |  |  |

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:

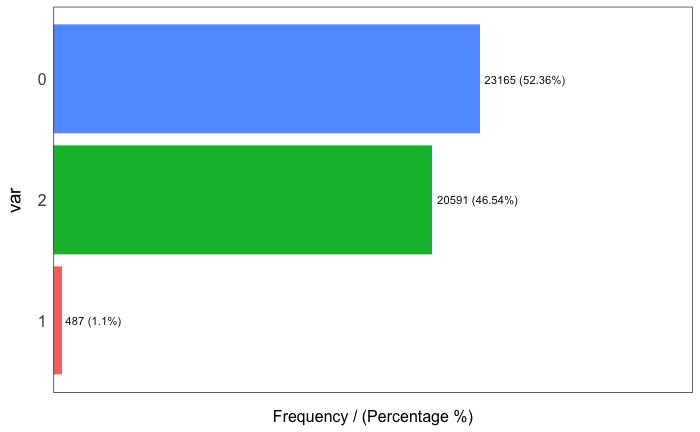
* CAT\_OCUP: 0 significa que la persona encuestada está desocupada y nunca tuvo una ocupación anterior.



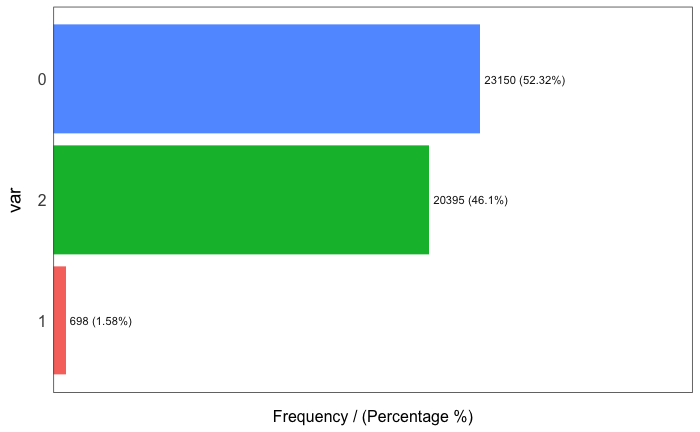
* CAT\_INAC: 0 significa que la persona tiene ocupación



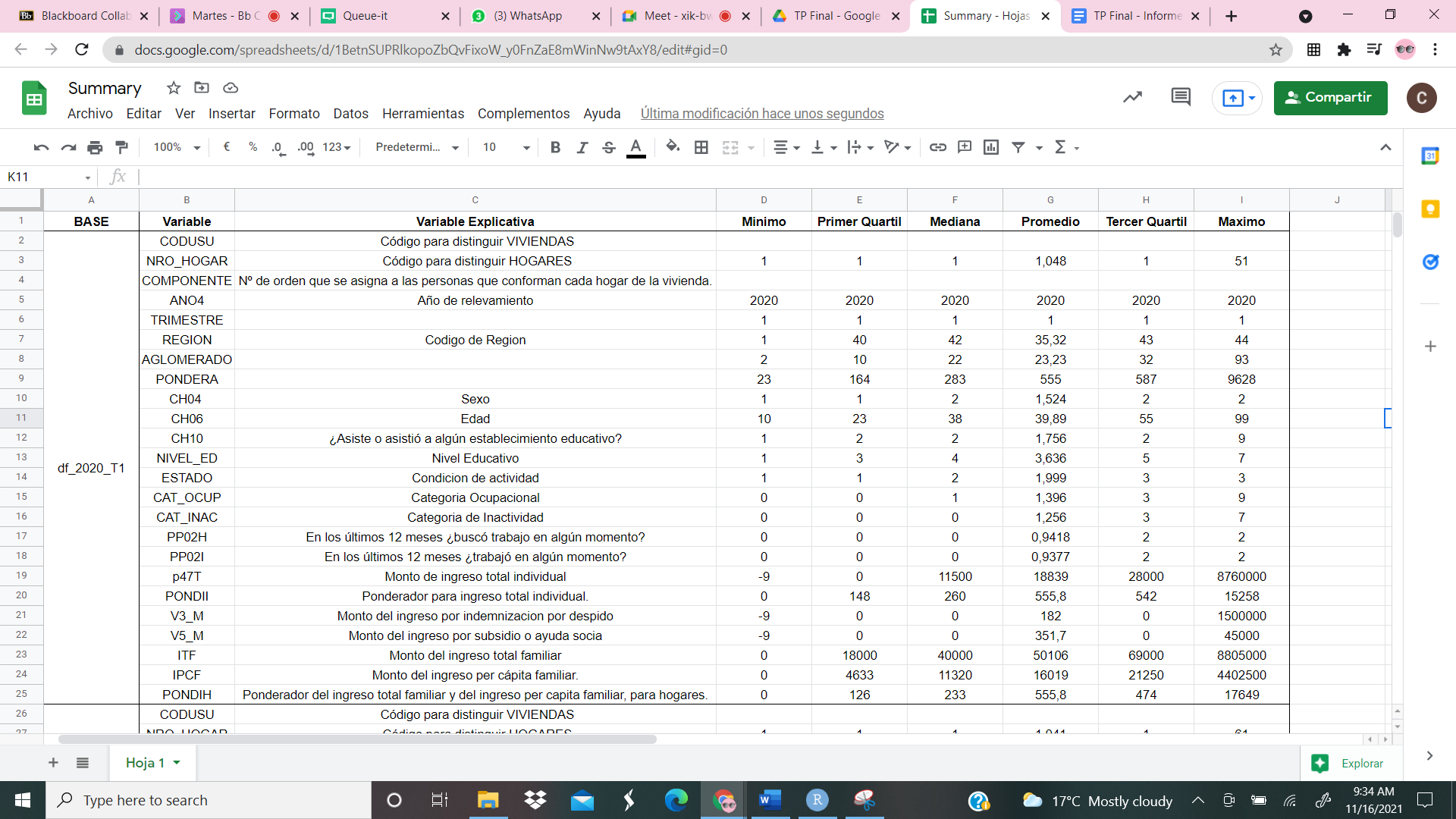
* PP02H: 0 significa que la persona tiene ocupación

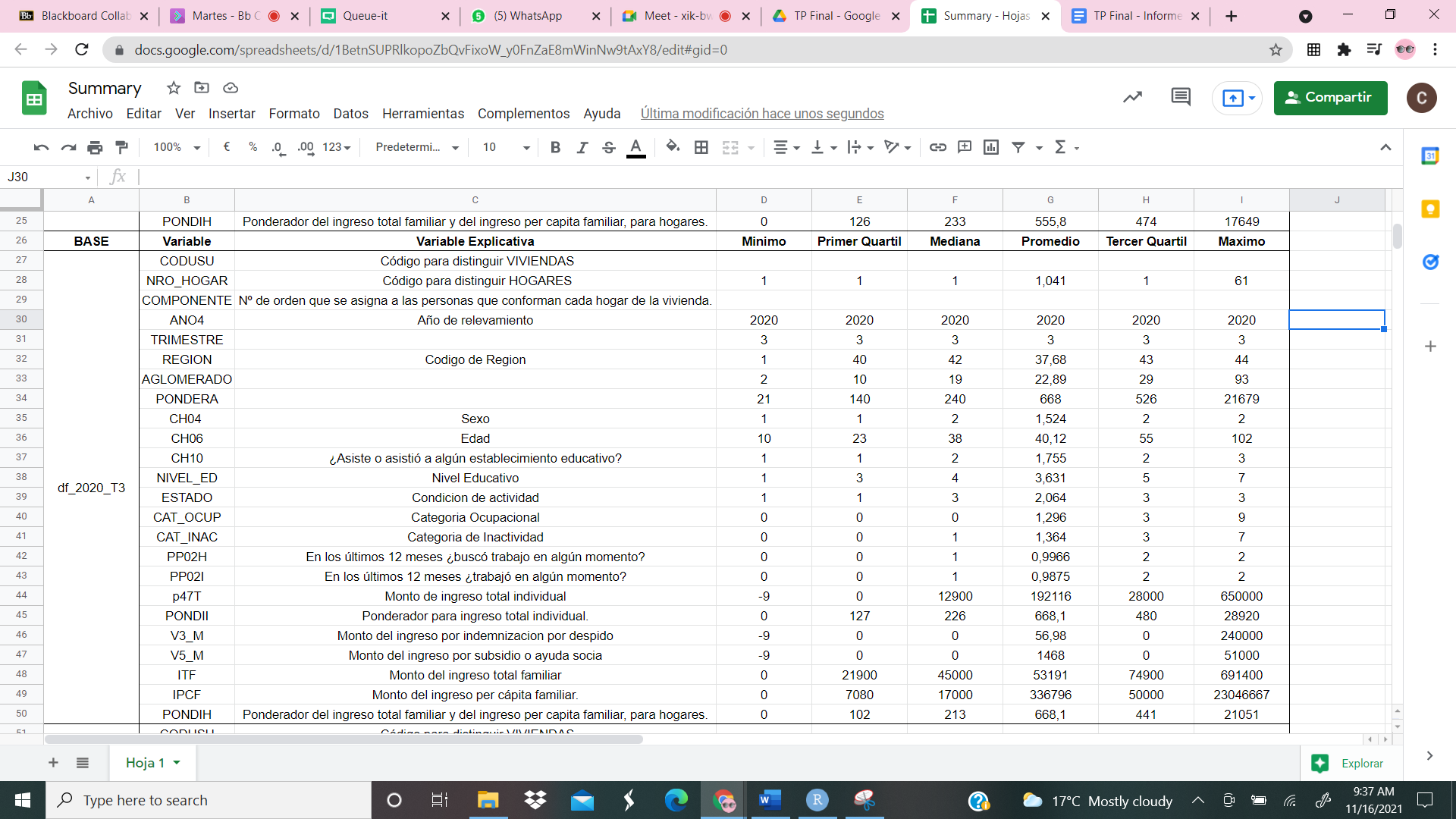


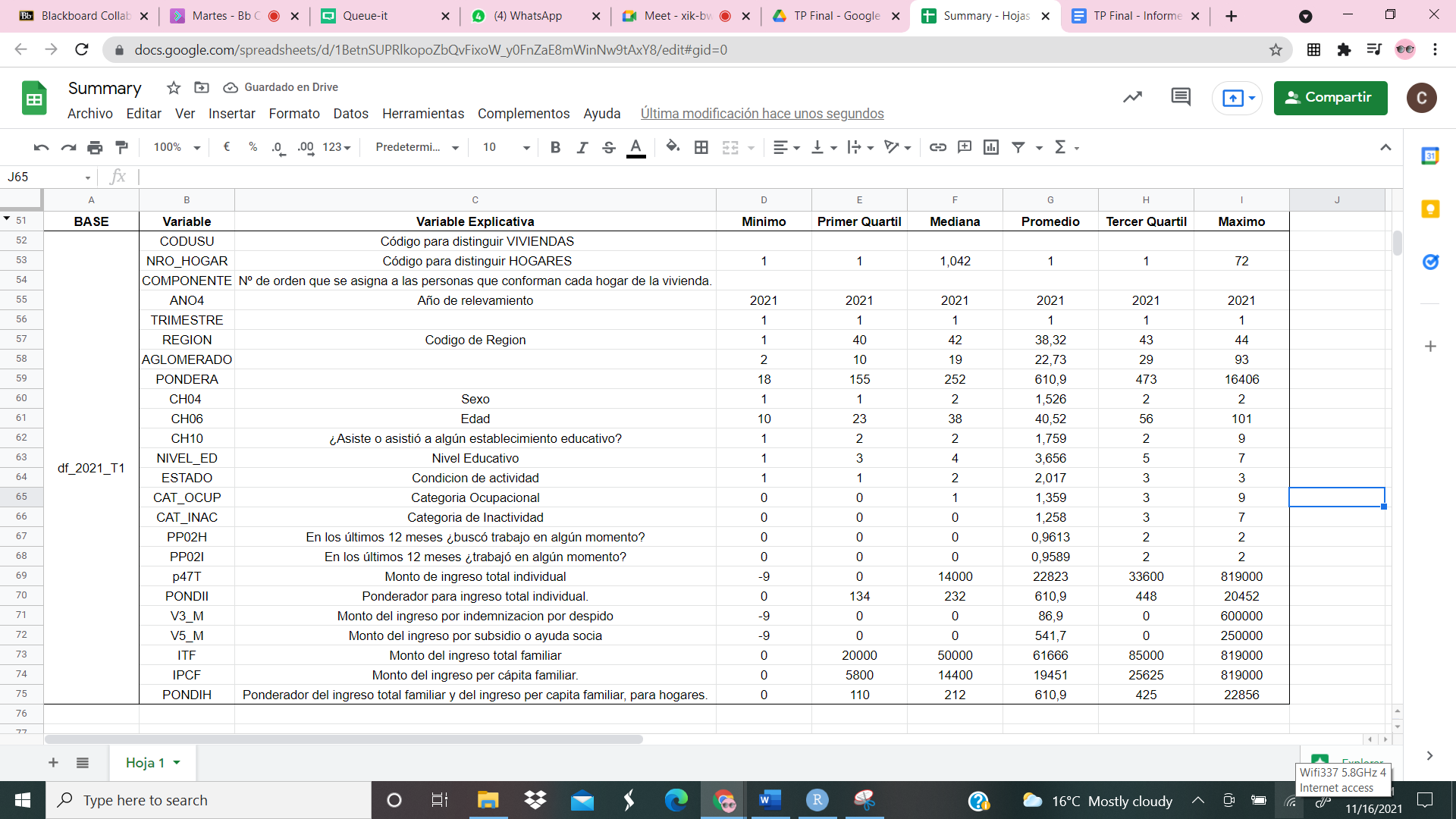
* PP02I: 0 significa que la persona tiene ocupación hace más de 12 meses



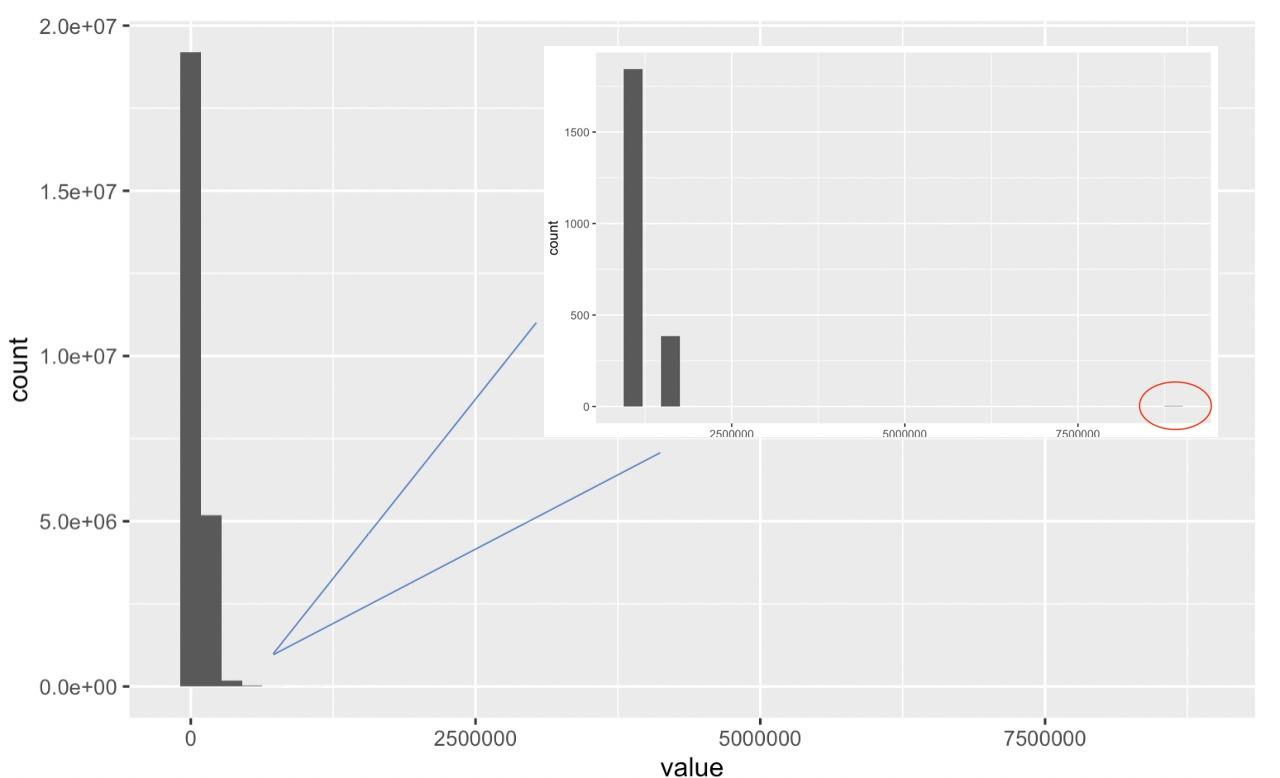
**SUMMARY**



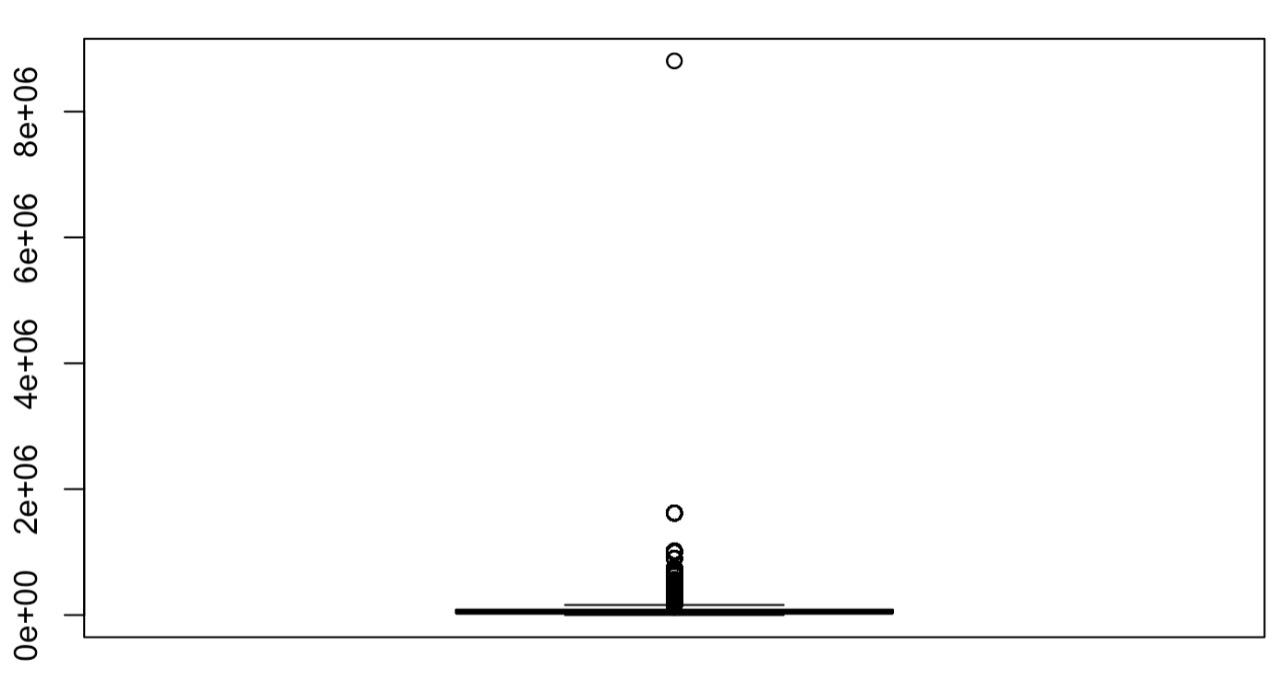




Para estudiar valores extremos separamos las variables categóricas de las cuantitativas. La mayoría de las variables son categóricas, por lo que no se puede hacer análisis de outliers sobre ellas. Las variables existentes cuantitativas son montos de ingreso (P47T, V3\_M, V5\_M, ITF y IPCF). Estas variables 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:



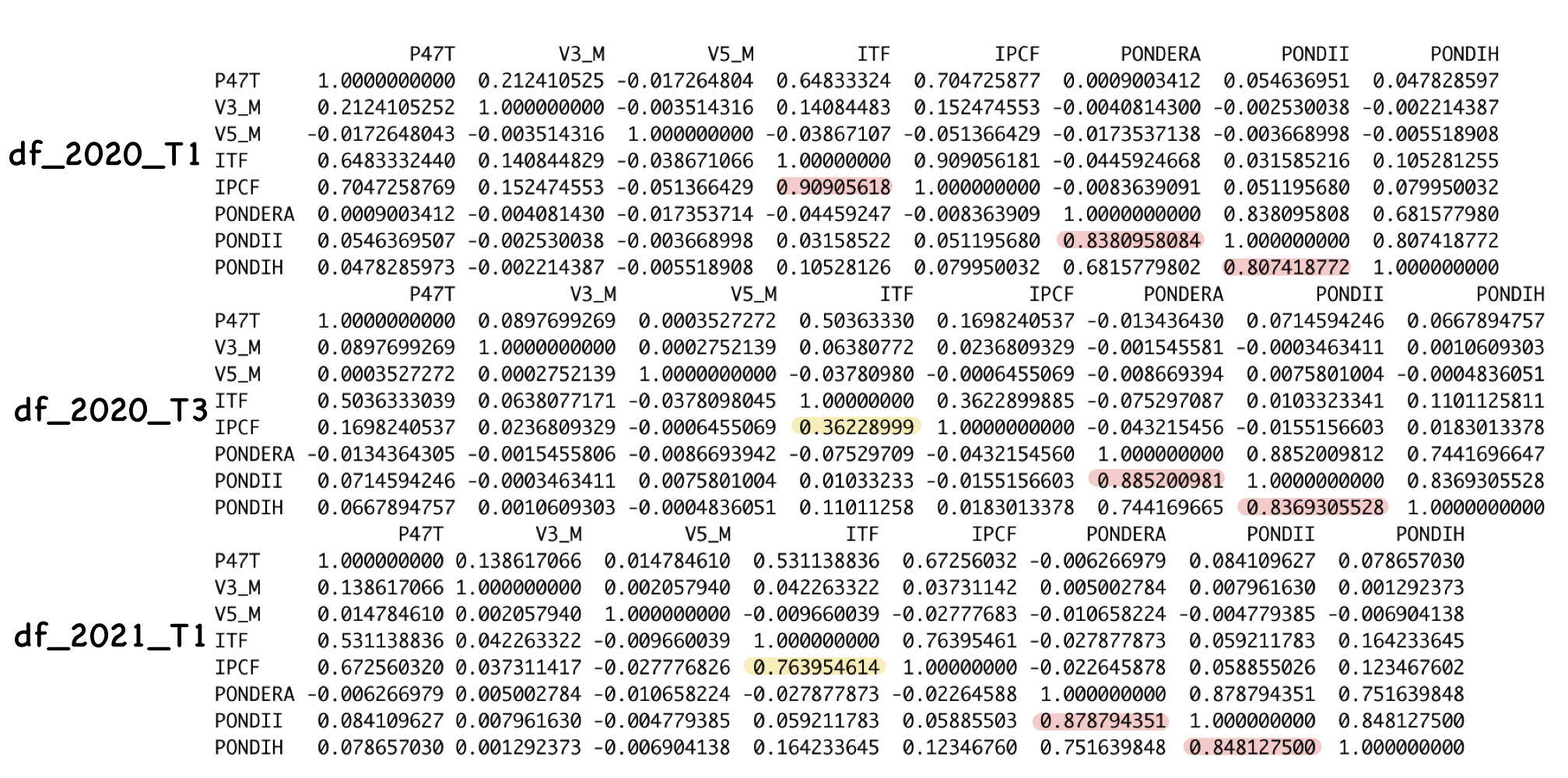
Por esto, si estudiamos los valores extremos encontraríamos muchos de ellos ya que los valores máximos tienden a ser muy altos y alejados del tercer cuartil.



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.



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 periodos, 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 trabajo
* 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](https://towardsdatascience.com/exploratory-factor-analysis-in-r-e31b0015f224#:~:text=KMO,a%20factor%20analysis.). Según este criterio el KMO debería ser mayor a 0.6 y los resultados en cada dataset fueron:

| **2020 T1** | **2020 T3** | **2021 T1** |
| --- | --- | --- |
| 0.53 | 0.52 | 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**

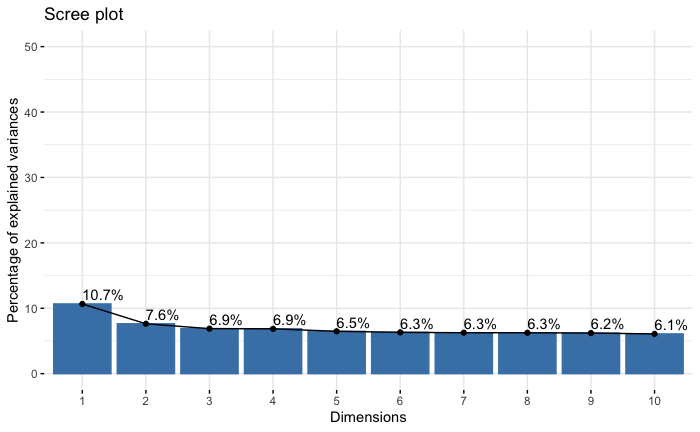
Se quiere estudiar sí las características qué describen a las personas qué toman planes son las mismas en períodos previos y durante la pandemia. Se estudiará sobre los datos de la muestra, sin ponderación por lo qué no se puede inferir qué las conclusiones suceden en la realidad de la población.

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.

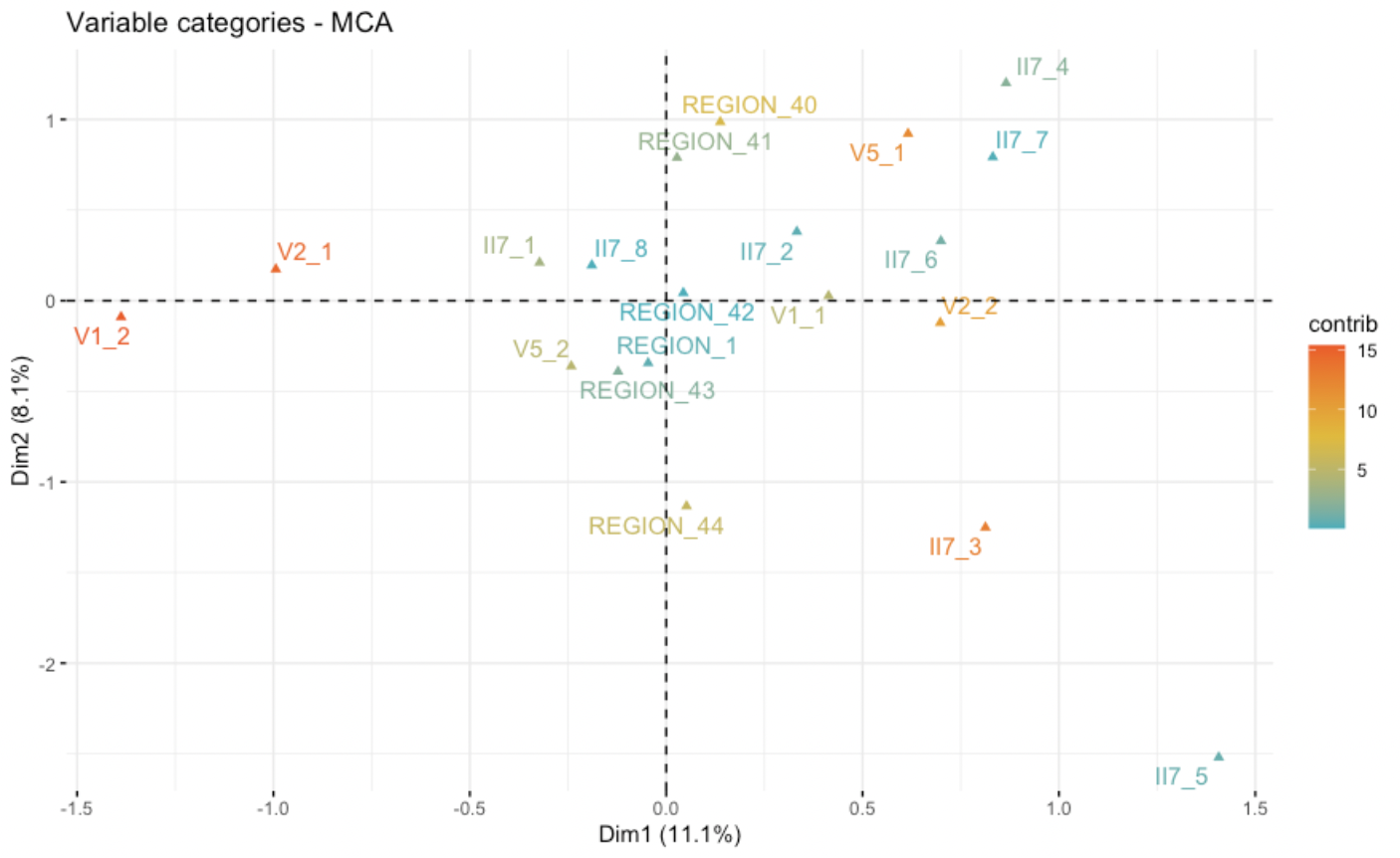
**% de varianza explicada**



**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

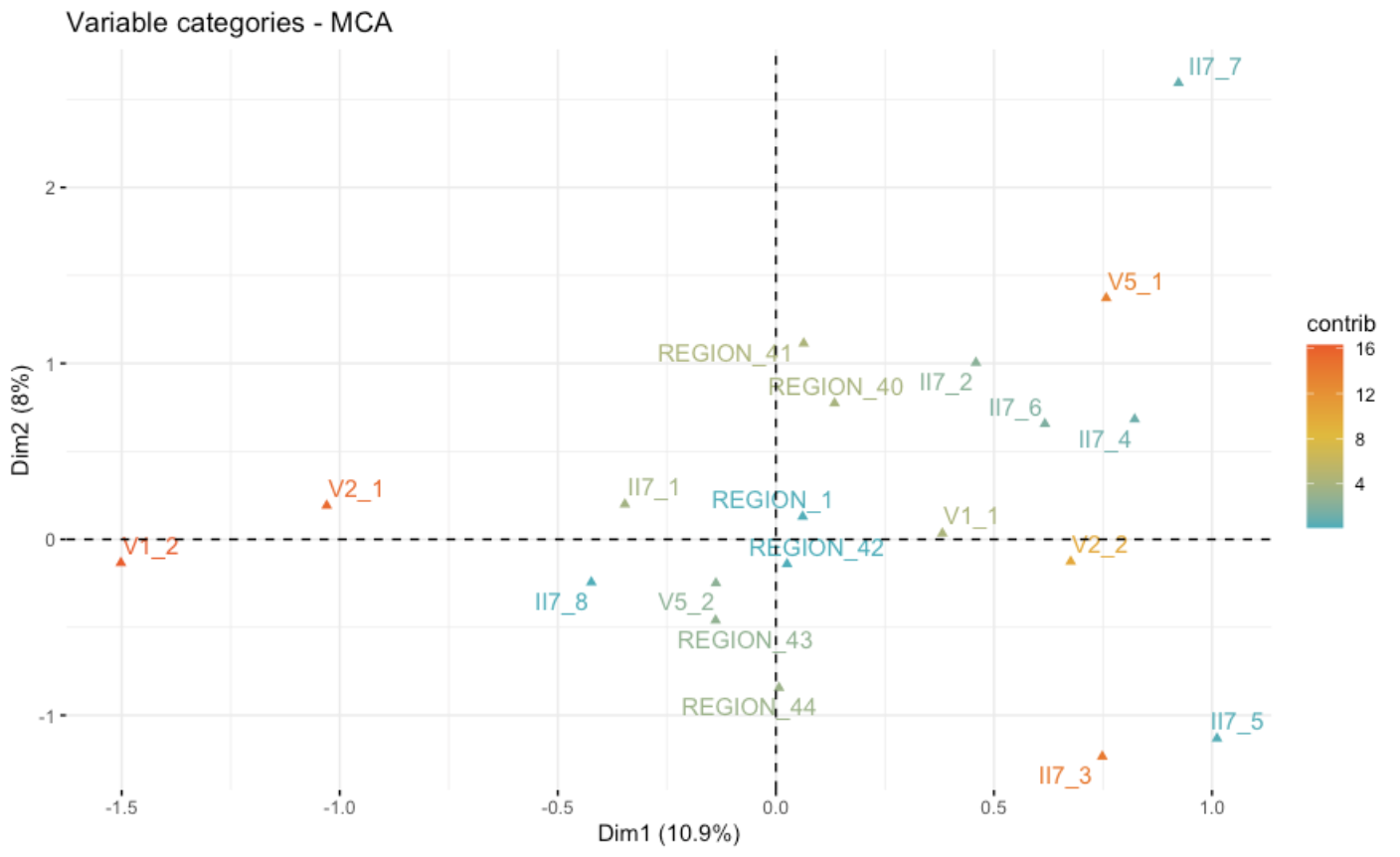


Podemos decir que las personas con alta Dim 1, son el grupo más afectado económicamente por la pandemia, a quienes se apuntaron muchos planes sociales como el IFE . Esto incluye a personas que trabajan en relación de dependencia, que alquilan y no disponen de vivienda propia, además cobran planes sociales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

**Análisis general pre-pandemia**



Análisis MCA

1. **Nuevos sectores necesitan ayudas sociales**: Acercamiento 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\*:** Acercamiento 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:***

X1 = Monto de ingreso por subsidio (V5\_M) en el primer trimestre del 2020.

X2 = Monto de ingreso por subsidio (V5\_M) en el tercer trimestre del 2020.

***Hipótesis:***

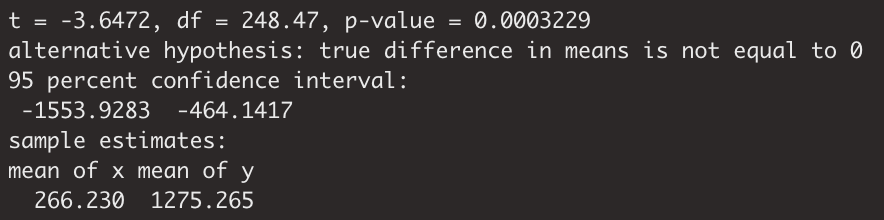
H0: mu(a) = mu(b)

H1: mu(a) ≠ mu(b)

***Condición de Rechazo:*** p-value ≤ 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 H0, se ve una diferencia significativa entre los montos de ingreso del primer y tercer trimestre.



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:***

X2 = Monto de ingreso por subsidio (V5\_M) en el tercer trimestre del 2020.

X3 = Monto de ingreso por subsidio (V5\_M) en el primer trimestre del 2021.

***Hipótesis:***

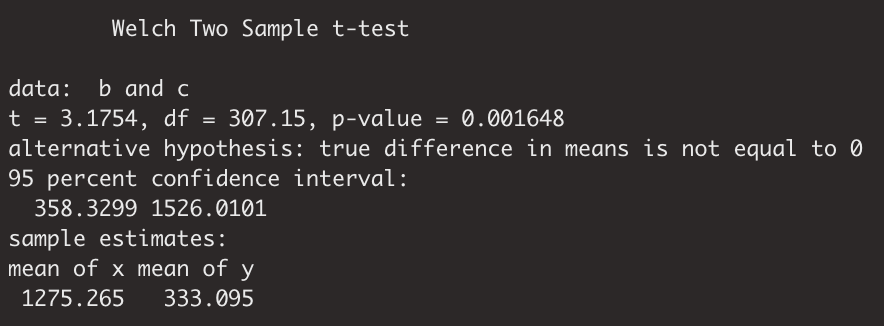
H0: mu(b) = mu(c)

H1: mu(b) ≠ mu(c)

***Condición de Rechazo:*** p-value ≤ 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 H0, se ve una diferencia significativa entre los montos de ingreso del tercer trimestre del 2020 y el primer trimestre del 2021.



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 periodos.

**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 sí acompaña el análisis anterior. Se debería encontrar crecimiento entre el primer trimestre y el tercer trimestre del 2020, ya qué 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:***

p2020 : proporción de personas que reciben planes sociales en el 1º trimestre de 2020 (V5=1)

p2021: proporción de personas que reciben planes sociales en el 1º trimestre de 2021 (V5=1)

***Hipótesis:***

H0: Durante la pandemia menos o igual cantidad de población recibió planes sociales (p2020 ≤ p2021)

H1: Durante la pandemia más gente recibió planes sociales (p2020 > p2021)

***Condición de Rechazo:*** p-value ≤ 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.