CoderHouse

Curso Data Science

Informe del Proyecto Final

Análisis socioeducativo de los habitantes de la Ciudad de Buenos Aires

Profesor: Damian Dapueto
Tutor: Héctor Alonso

Grupo de Trabajo:

Lucia Buzzeo, Lucia Hukovsky,
Jose Saint German, Juan Martín Carini

5 de septiembre de 2022

Índice

1. Introducción

1.1. Resumen del proyecto

La ciudadanía es un concepto jurídico, filosófico y político que ha sido creado para designar a una persona física que constituye una sociedad o entidad territorial. Para las personas que forman parte de una comunidad, ciudadanos, resulta de suma importancia sentirse representados por los demás integrantes de la misma, mediante políticas públicas que abarquen sus necesidades y requerimientos.

La toma de datos demográficos y la estadística son dos herramientas primordiales a la hora de identificar requerimientos de los integrantes de una comunidad. Dichas herramientas describen, de forma cuantitativa, a la sociedad bajo estudio. Precisamente, los censos y la estadística son la fuente primaria de información para la planificación económica y social de una población, por parte de sus representantes.

En el caso particular de Argentina, el Instituto Nacional de Estadística y Censos (INDEC) es el organismo público que ejerce la dirección superior de todas las actividades estadísticas oficiales que se realizan en el país. La información que produce el INDEC es una herramienta básica para la planificación de políticas públicas, así como para las investigaciones y proyecciones que se realizan en los ámbitos académico y privado.

Al adentrarse y estudiar los índices correspondientes a uno de los ejes principales, educación, en un territorio delimitado, Ciudad Autónoma de Buenos Aires, se ha encontrado una gran limitación relacionada con su acceso inequitativo para los diferentes actores de la sociedad.

Este hecho tiene consecuencias de índole social y económico para la población. Sin embargo, la principal problemática se da a nivel individual, y radica en el impedimento al acceso educativo para un porcentaje de la sociedad. Esto no ha resultado una novedad para el grupo, pero sí ha dado el pie a la búsqueda de una respuesta teórica a dicha disparidad, en concreto, a descubrir las principales variables que afectan el nivel educativo.

El análisis realizado en el marco del presente proyecto podría establecer una base de requerimientos que permitan generar políticas públicas efectivas, no solo en el ámbito educativo, sino en el económico, cultural, social y geográfico, entre otros.

1.2. Objetivos del proyecto

En el presente proyecto se persigue el objetivo de descubrir las principales variables intervinientes en el nivel máximo educativo alcanzado por la población de la Ciudad Autónoma de Buenos Aires (CABA).

En concreto, con este proyecto se plantea el análisis de variables que podrían contribuir al nivel máximo educativo alcanzado por cada individuo en CABA. Las mismas se estudian tanto de forma independiente como entrelazadas. El alcance de dicho objetivo permitiría identificar posibles causas que derivan en la desigualdad de acceso educativo, y, en última instancia, generar políticas públicas eficaces que permitan subsanar dicha problemática.

De este objetivo principal se desprenden los siguientes objetivos específicos:

- Determinar si la ubicación geográfica del encuestado es determinante para alcanzar ciertos niveles educativos. De este objetivo se desprende determinar la relación entre el nivel educativo y la comuna del encuestado, así como la relación entre la misma variable y el hecho de que el encuestado habite en una villa de emergencia.
- Establecer la fuerza con la que el nivel socioeconómico afecta la variable target.
- Explorar la relación del target con otras variables, como el sexo del encuestado, la cantidad de hijos, la afiliación de salud o la edad.

1.3. Definición de la fuente de información

Para trabajar esta problemática, se ha recurrido a la Encuesta Anual de Hogares del Gobierno de la Ciudad de Buenos Aires para el año 2019. El dataset está disponible en la base de datos abiertos del GCBA.

Esta encuesta contiene información demográfica, social, económica, educativa y de salud de 14319 habitantes de la Ciudad, la cual es una muestra representativa que permite obtener un vistazo de la población de la Ciudad.

2. Planificación

Estructura de los trabajos

Este trabajo se ha dividido en 3 partes:

1. **Introducción a las variables del problema:** Se realiza un análisis de las variables del dataset. En el mismo se busca conocer su performance dentro del dataset. A la vez, se investiga cómo las variables interactúan entre sí. Esta parte es lo que se conoce como análisis univariado, bivariado y multivariado

2. **Modelos análiticos:** En esta sección se llevan a cabo diversos modelos análiticos y algoritmos que sirven para alcanzar los objetivos seteados para el presente proyecto. Como la variable objetivo es categórica, se realizan diversos modelos de clasificación.

3. **Conclusión:** Se alcanzan conclusiones finales sobre los hallazgos. Además, se discuten posibles limitaciones y se plantean futuras líneas de análisis, a partir del análisis presente.

3. Introducción a las variables: Análisis exploratorio de los datos

Una ves cargado el dataset con el que vamos a trabajar, miramos sus variable, el tipo que son y si tienen nulls:

RangeIndex: 14319 entries, 0 to 14318
Data columns (total 31 columns):

#	Column	Non-Null Count	Dtype
0	id	14319 non-null	int64
1	nhogar	14319 non-null	int64
2	miembro	14319 non-null	int64
3	comuna	14319 non-null	int64
4	dominio	14319 non-null	object
5	edad	14319 non-null	int64
6	sexo	14319 non-null	object
7	parentesco_jefe	14319 non-null	object
8	situacion_conyugal	14318 non-null	object
9	num_miembro_padre	14319 non-null	object
10	num_miembro_madre	14319 non-null	object
11	estado_ocupacional	14319 non-null	object
12	cat_ocupacional	14319 non-null	object
13	calidad_ingresos_lab	14319 non-null	object
14	ingreso_total_lab	14319 non-null	int64
15	calidad_ingresos_no_lab	14319 non-null	object
16	ingreso_total_no_lab	14319 non-null	int64
17	calidad_ingresos_totales	14319 non-null	object
18	ingresos_totales	14319 non-null	int64
19	calidad_ingresos_familiares	14319 non-null	object
20	ingresos_familiares	14319 non-null	int64
21	ingreso_per_capita_familiar	14319 non-null	int64
22	estado_educativo	14319 non-null	object
23	sector_educativo	14316 non-null	object
24	nivel_actual	14319 non-null	object
25	nivel_max_educativo	13265 non-null	object
26	años_escolaridad	14257 non-null	object
27	lugar_nacimiento	14318 non-null	object
28	afiliacion_salud	14315 non-null	object
29	hijos_nacidos_vivos	6535 non-null	object
30	cantidad_hijos_nac_vivos	14319 non-null	object

dtypes: int64(10), object(21)
memory usage: 3.4+ MB

Ahora, en base a los datos arrojados por la tabla de arriba, generamos diversas transformaciones de variables, así como la creación de la variable "Target", pues es la que usaremos para todo el análisis:

- Creamos la variable "Target" y le asignamos la varibale "nivel_max_educativo".
- En la variable "Target", reducimos su dimensionalidad intercambiando los valores:
  - "Secundario/medio comun" y "EGB (1° a 9° año)" por "sec_completo",
  - "Primario especial" y "Primario comun" por "prim_completo",
  - "Sala de 5" por "incial",
  - "Otras escuelas especiales"por "superior",
  - y por último a "No corresponde" por nulos.

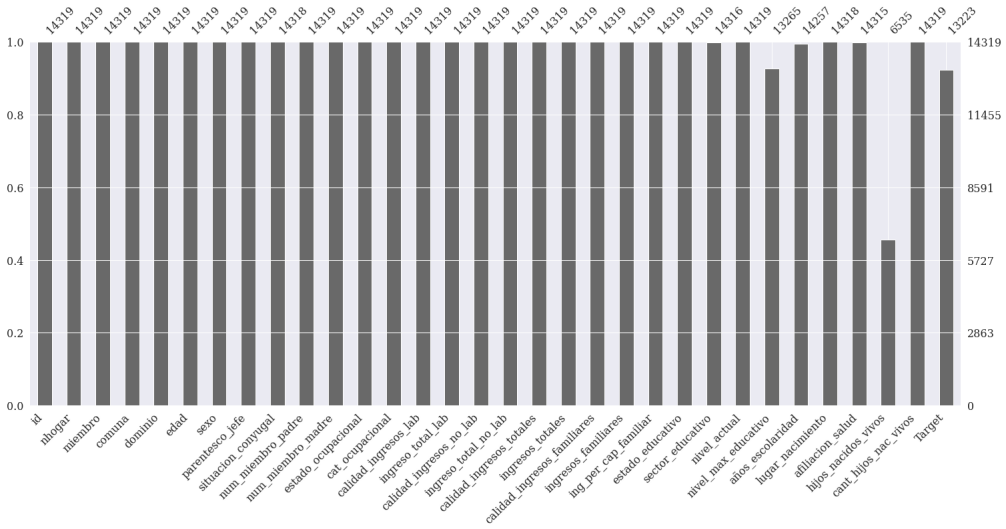
- Remplazamos los valores de “años_escolaridad” para que todos sean numéricos.
- En la variable “cantidad_hijos_nac_vivos” cambiamos el valor “no corresponde” como nulo, para luego cambiar el tipo de variable a entero.
- Las variables “comuna”, “id”, “nhogar” y “miembro” son de tipo numérico, pero deberían ser categóricas, por lo tanto transformamos su tipo a string.
- Por último renombramos algunas variables para que sean más cortas:
  - “dominio_Villas_de_emergencia” por “dominio_villas”,
  - “ingreso_per_capita_familiar”por “ing_per_cap_familiar”,
  - “cantidad_hijos_nac_vivos” por “cant_hijos_nac_vivos”.

A continuación detallamos un diccionario de las variables:

- **“id”** : Clave que identifica la vivienda,
- **“nhogar”** : La variable id + nhogar = clave que identifica a cad hogar”,
- **“miembro”** : Variables id + nhogar + miembro = clave que identifica cada persona”,
- **“comuna”** : Comuna donde reside la persona encuestada,
- **“edad”** : Edad de la persona encuestada,
- **“sexo”** : Sexo de la persona encuestada,
- **“parentesco_jefe”** : Parentesco entre la persona encuestada y el jefe d” hoga,
- **“situacion_conyugal”** : Situación conyugal de la persona encuestada,
- **“num_miembro_padre”** : Número de miembro que corresponde al padre,
- **“num_miembro_madre”** : Número de miembro que corresponde a la madre,
- **“estado_ocupacional”** : Situación ocupacional de la persona encuestada,
- **“cat_ocupacional”** : Categoría ocupacional de la persona encuestada,
- **“calidad_ingresos_lab”** : Calidad de la declaración de ingresos laborales totales,
- **“ingreso_total_lab”** : Ingreso total laboral percibido el mes anterior,
- **“calidad_ingresos_no_lab”** : Calidad de la declaración de ingresos no laborale totales”,
- **“ingreso_total_no_lab”** : Ingreso total no laboral percibido el mes anterior,
- **“calidad_ingresos_totales”** : Calidad de ingresos totales individuales,
- **“ingresos_totales”** : Ingreso total individual percibido el mes anterior,
- **“calidad_ingresos_familiares”** : Calidad de ingresos totales familiares,
- **“ingresos_familiares”** : Ingresos totales familiares percibido el mes anterior,
- **“ing_per_cap_familiar”** : Ingreso familiar per capita percibido el mes anterior,
- **“estado_educativo”** : Asistencia (pasada o presente) o no a algú establecimiento educativo”,
- **“sector_educativo”** : Sector al que pertenece el establecimiento educativo a que asiste”,
- **“nivel_actual”** : Nivel cursado al momento de la encuesta,
- **“nivel_max_educativo”** : Máximo nivel educativo que se cursó,
- **“años_escolaridad”** : Años de escolaridad alcanzados,
- **“lugar_nacimiento”** : Lugar de nacimiento de la persona encuestada,
- **“afiliacion_salud”** : Afiliación de salud de la persona encuestada,
- **“hijos_nacidos_vivos”** : Tiene o tuvo hijos nacidos vivos,

- **"cant_hijos_nac_vivos"** : Cantidad de hijos nacidos vivos,
- **"dominio"** : ¿la vivienda se ubica en una villa de emergencia?,
- **"Target"** : Nivel máximo educativo.

Y comenzamos el analisis EDA del mismo, primero analizando los nulos:

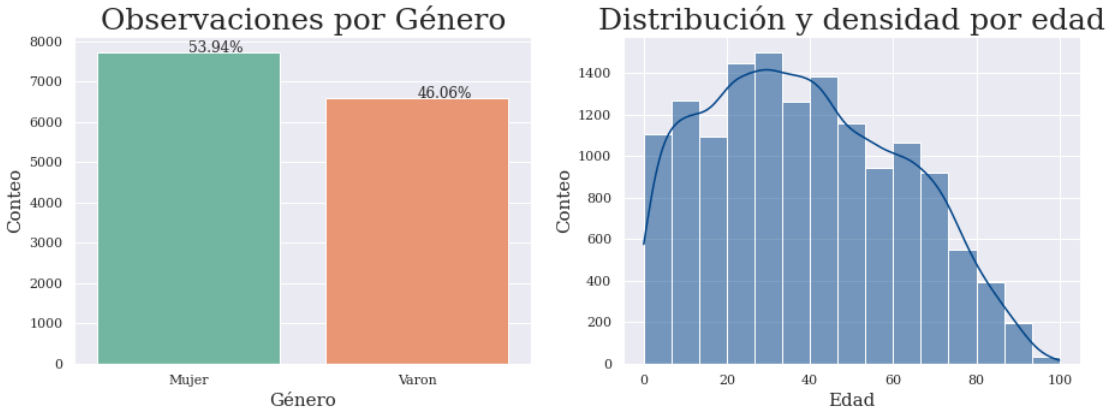


Así, detectamos que nuestra variable target tiene 1054 valores nulos. Es importante tener este dato presente cuando queramos correr un algoritmo de clasificación.

3.1. Análisis univariado

3.1.1. Género y edad

Comenzamos con un pantallazo general sobre las primeras cualidades de los datos, como muestra representativa para la EPH, sobre quiénes son los ciudadanos representados en el dataset.

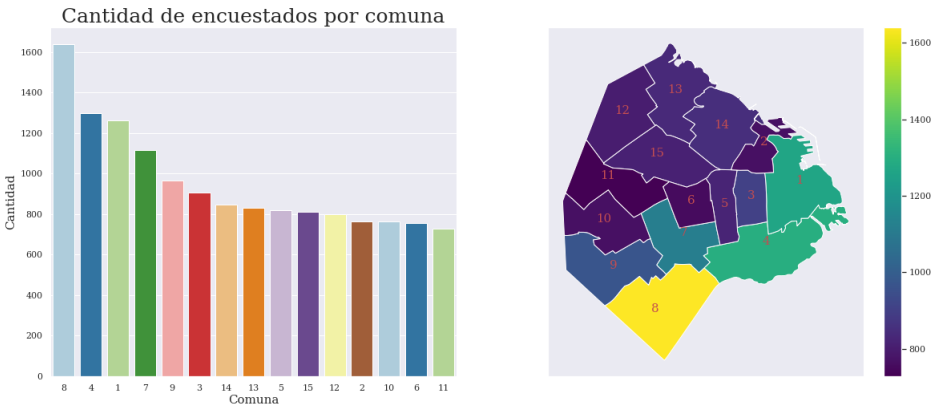


En la variable género los datos parecen equilibrados en las categorías. Para el caso de la variable "edad", la distribución se asemeja a la de una normal.

3.1.2. Comuna

Seguimos observando la variable "comuna". En la misma se muestra la comuna de la Ciudad de Buenos Aires del entrevistado, de manera de tener una ubicación geográfica. Consideramos importante revisar esta variable ya que tenemos como hipótesis que el nivel educativo alcanzado puede estar dependiendo de la zona geográfica de la ciudad en la que se encuentra el entrevistado.

Para esto vamos a generar un mapa, así que utilizaremos el mapa de comunas de la Ciudad de Buenos Aires, transformamos las variables que vamos a usar para joinear el mapa con la base de manera que coincidan, transformamos la base para contabilizar la frecuencia con la que aparece cada comuna en la base. Y por último unimos ambos datasets y generamos una nueva variable con las coordenadas para poder agregar etiquetas en el centro geográfico de cada comuna, que nos da como resultado los siguientes gráficos:

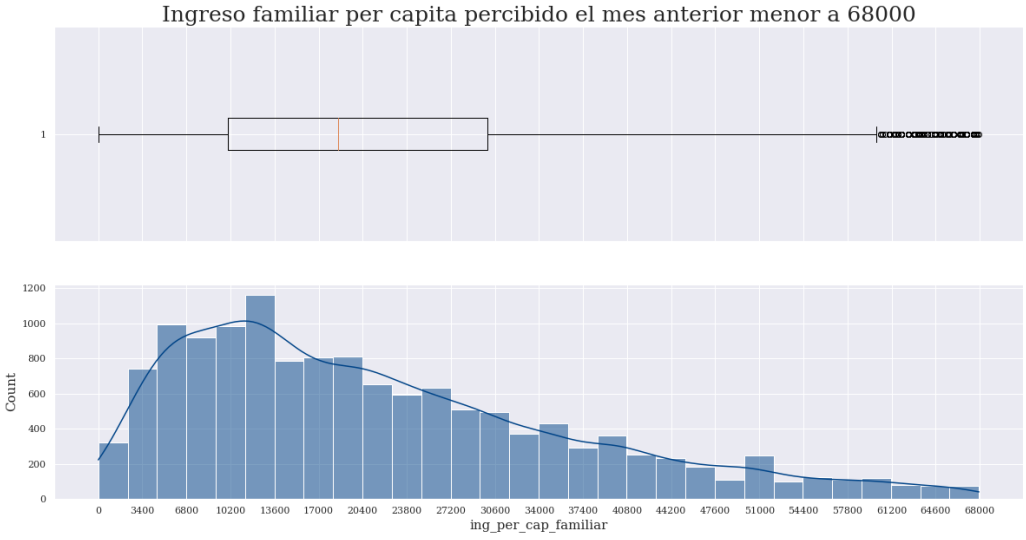


Observando ambos gráficos vemos que las comunas 1,4,7 y 8 tienen mayor cantidad de casos. Queda por verse si en posteriores análisis es necesario abordar esta diferencia para evitar sesgos. Para eso, será necesario tomar en cuenta el porcentaje de la población total de cada comuna.

3.1.3. Ingreso familiar per cápita

Ahora probamos con observar los ingresos familiares. Creemos que puede ser un indicador interesante del nivel educativo.

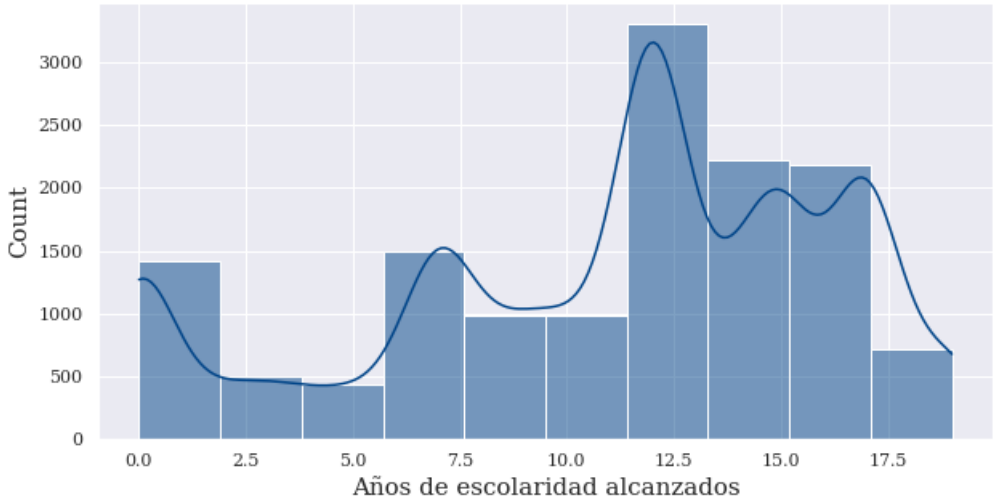
Para esto, armamos una función para graficar y jugar con el nivel del filtrado de la variable y obtener un histograma que permita apreciar mejor la distribución de la variable sin tantos outliers:



Y luego de remover los outliers la distribución de los ingresos familiares sigue estando **sesgada**.

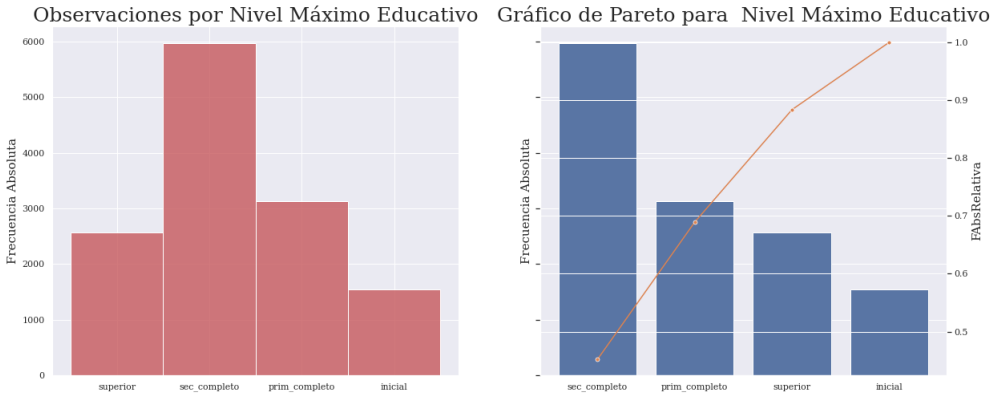
3.1.4. Años de escolaridad

Analizamos mediante un gráfico de barras los años de escolaridad alcanzados por los encuestados:



A simple vista se observan tres "picos": en el valor mínimo, alrededor del 7.5 y alrededor del 12.5. Podemos inferir que estos tres casos corresponden a no tener estudios, solo haber transcurrido el primario y haber transcurrido hasta la educación secundaria, respectivamente.

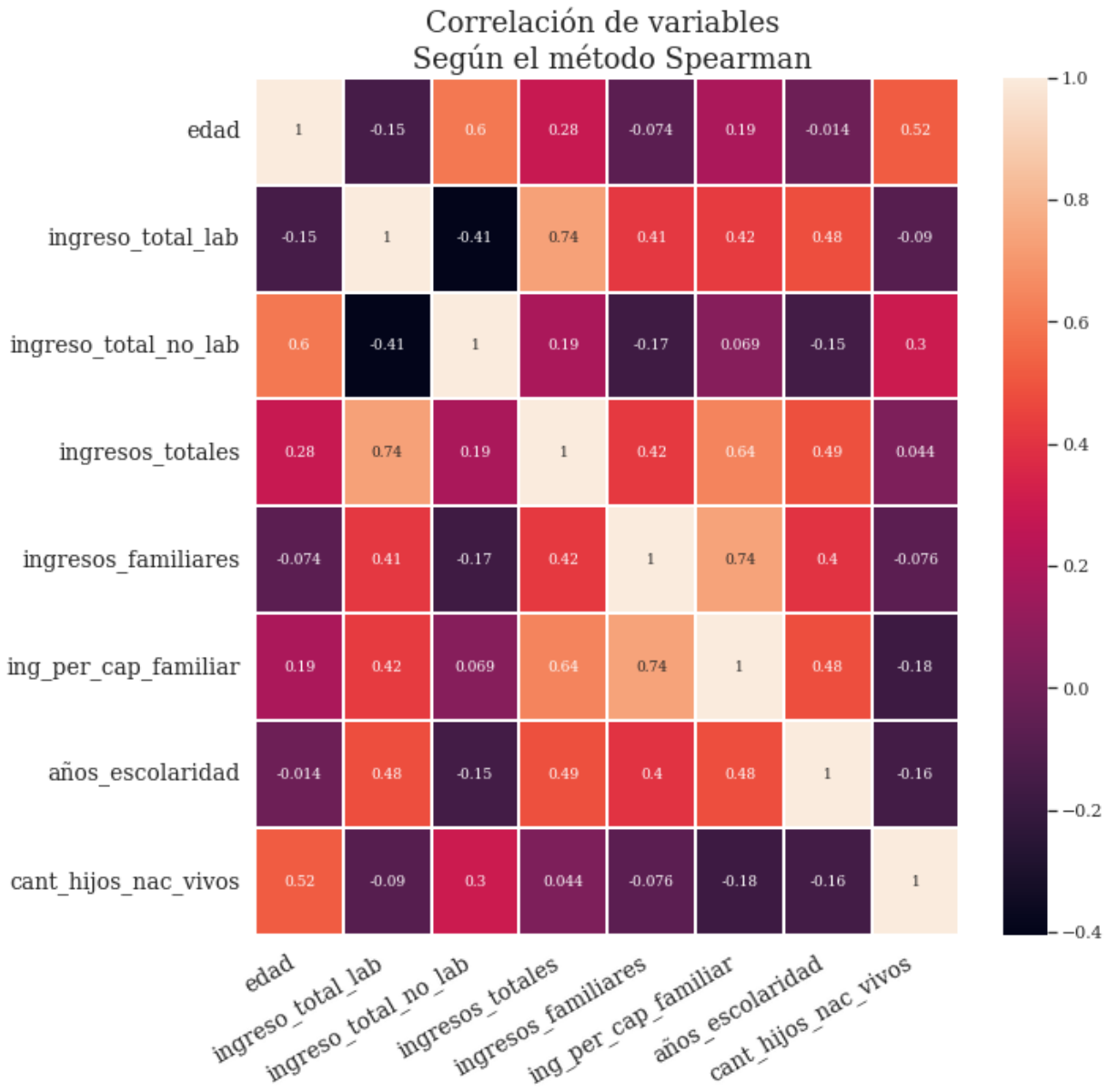
3.1.5. Máximo nivel educativo (Target)



Podemos observar que el nivel máximo educativo más alcanzado es el secundario completo, seguido por el primario. Contrario de lo que habíamos intuido anteriormente, el nivel superior quedó en tercer lugar. Adicionalmente, el nivel secundario y primario explican casi el 77 % de los datos.

3.2. Análisis bivariado

Para comenzar el análisis bivariado del problema, realizamos diferentes heat maps para ver si algo nos llama la atención entre las variables numéricas.

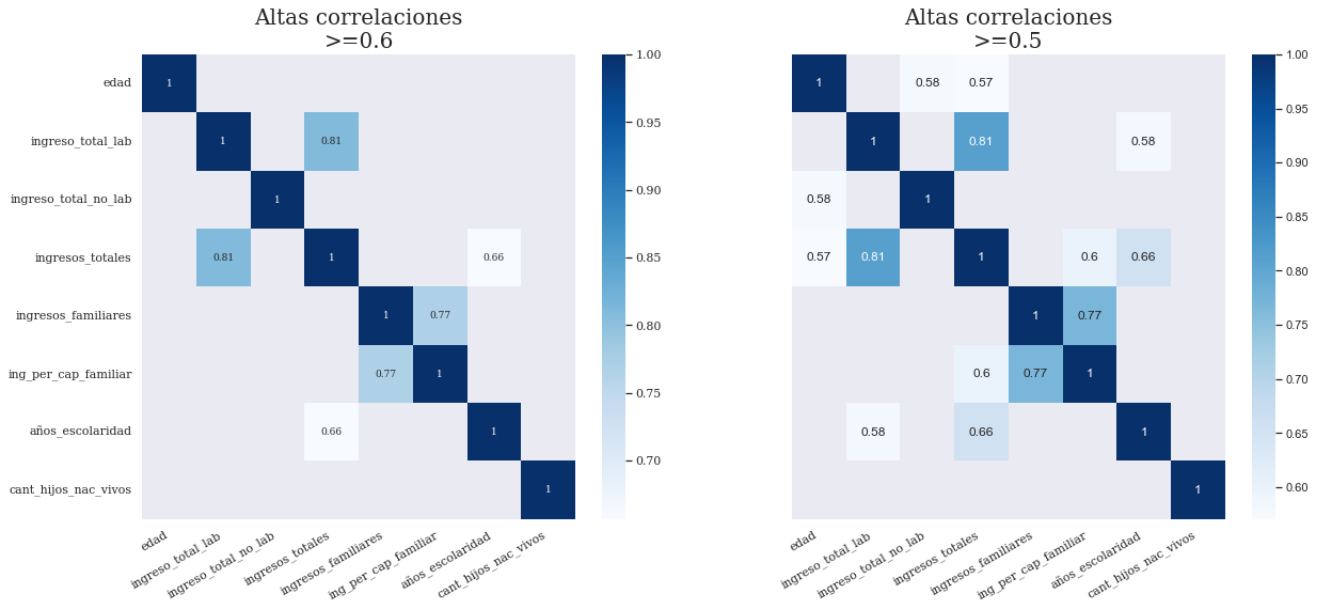


A simple vista, no se observan fuertes correlaciones.

Podemos notar que la variable "años_escolaridad" correlaciona moderadamente bien con variablres relacionadas al ingreso.

La principal correlación positiva es "años_escolaridad" con ingreso familiar per cápita ("ing_per_cap_familiar"), lo cual hace sentido teórico.

Ahora, utilizando el método para solo graficar en base a un threshold vemos que los años de escolaridad alcanzados por los entrevistados tienen algo relación (66 %) con la variable "ingresos_totales"



Por último corremos una tabla de correlación y filtramos las de valores más altos

	Variable_1	Variable_2	corr_value
2	ingreso_total_lab	ingresos_totales	0.80
6	ingresos_familiares	ing_per_cap_familiar	0.76
4	ingresos_totales	ing_per_cap_familiar	0.62
5	ingresos_totales	años_escolaridad	0.60
1	edad	ingreso_total_no_lab	0.57
7	años_escolaridad	Target	0.57
3	ingreso_total_lab	años_escolaridad	0.54

Conclusiones:

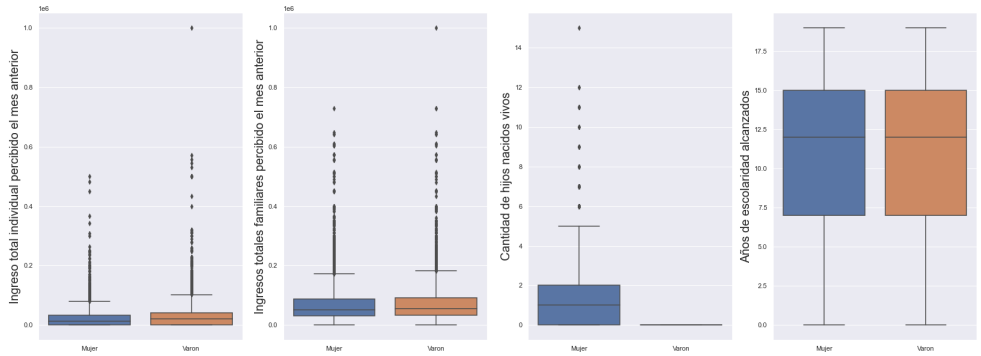
- Como es esperable, hay alta correlación entre las variables relacionadas al ingreso.
- A su vez, encontramos una alta correlación (66 %) entre los ingresos y los años de escolaridad.
- También observamos una relación positiva entre la edad y los ingresos totales.

3.2.1. Comparación entre variables numéricas

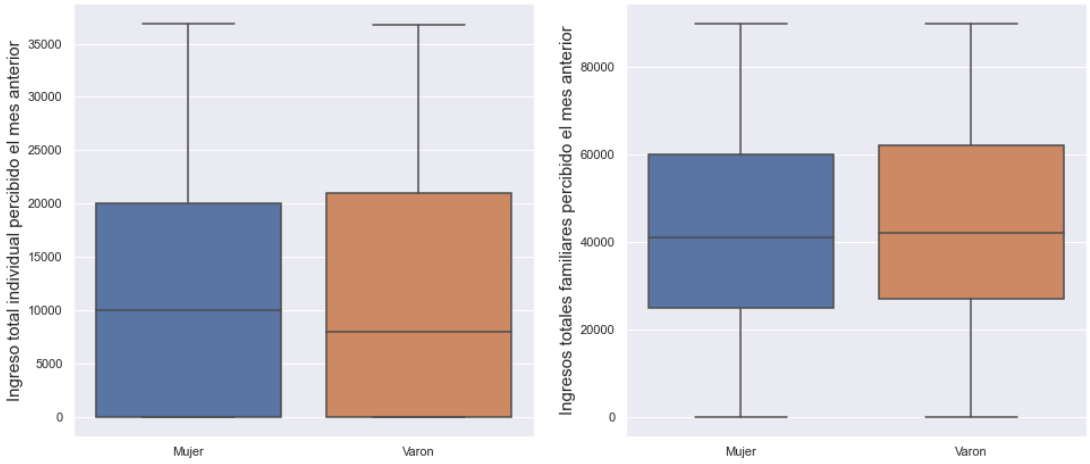
Se puede ver que desde los 30 años en adelante el ingreso total de la persona se corresponde con el ingreso familiar. Por ende suele haber un único ingreso fuerte por grupo familiar.

3.2.2. Comparación de variables categóricas con numéricas

Adicionalmente, vamos a comparar algunas variables con nuestro target, comenzando con los ingresos totales.

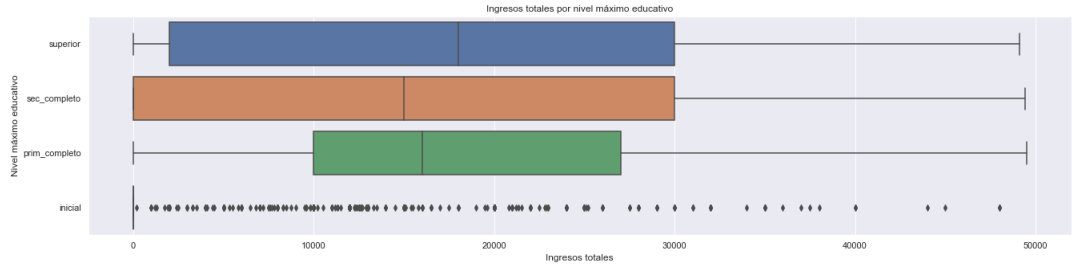


Probemos quitando outliers, a excepción de la cantidad de hijos nacidos vivos (puesto que no arrojará ningún dato nuevo) y de años de escolaridad (que no tiene outliers)

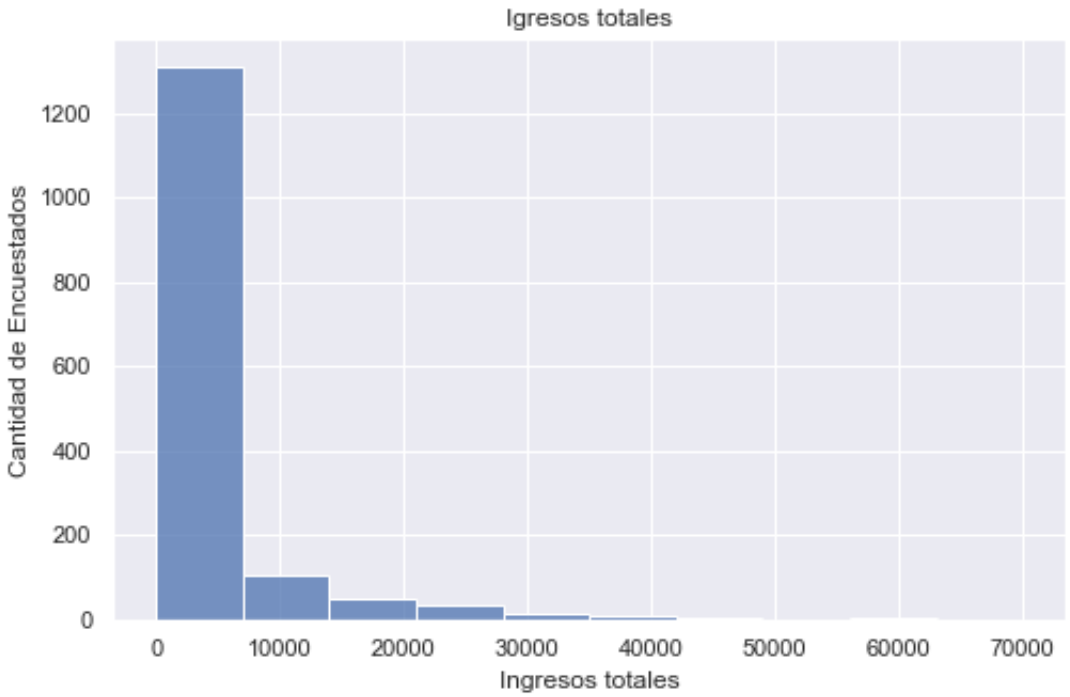


Por parte de las variables de ingreso, no parece haber nada disruptivo. La distribución por ingreso y años de escolaridad pareciera ocurrir pero no en un orden lineal.

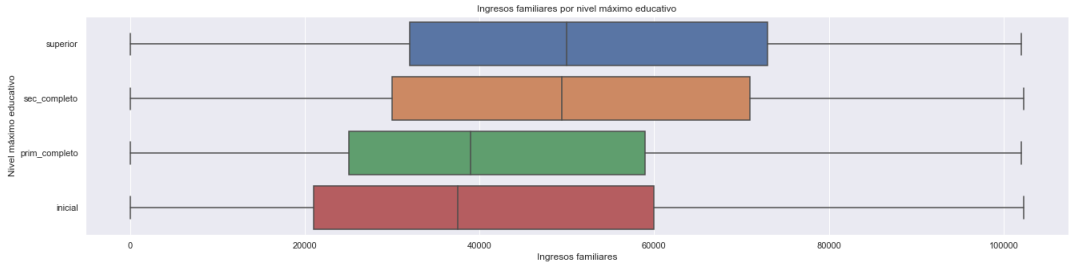
Llama la atención la variable "sexo": por algún motivo, todos los encuestados hombres figuran sin hijos nacidos vivos. Alternativamente, se podría investigar la metodología de la encuesta para ver si hay alguna respuesta. Adicionalmente, los hombres parecieran tener ingresos totales y familiares mayores que las mujeres, pero no pareciera que haya distribuciones desiguales en los años de escolaridad.



Parece que para el nivel inicial la remoción de outliers en otra categoría sigue siendo insuficiente para mostrar la distribución real de la variable. Echemos un vistazo a los valores de esta categoría.

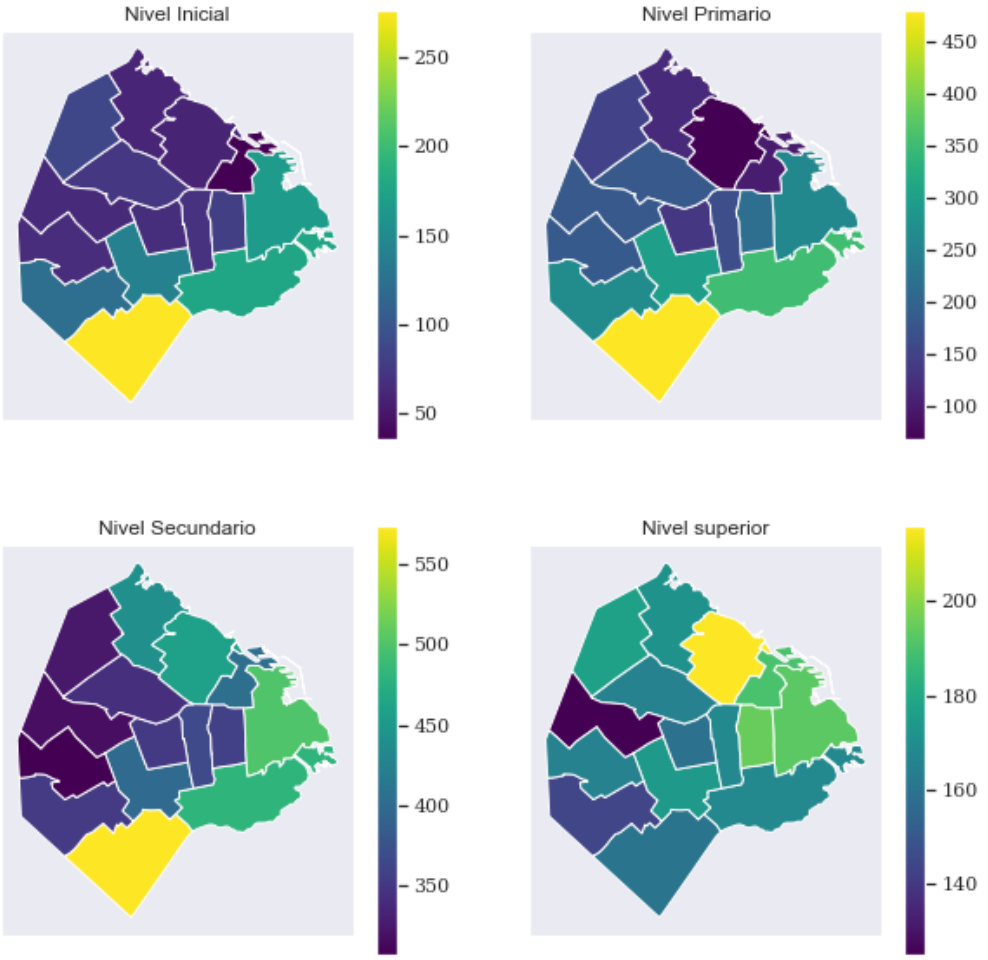


Lógicamente, la enorme mayoría de los ingresos tienen el valor inicial de 0, puesto que incluye a personas que en ese momento estaban cursando su educación inicial, por lo que tenían entre 2 y 6 años.

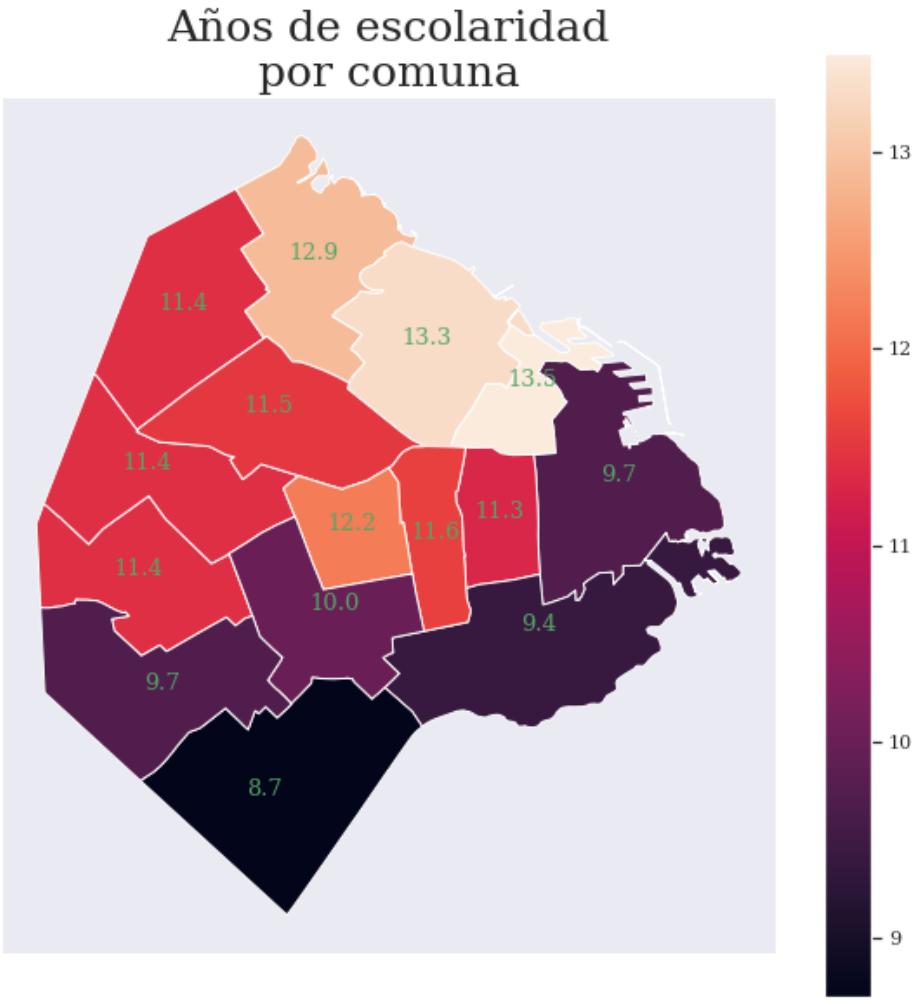


En definitiva, se observa un corrimiento de los valores centrales (dentro de la caja) hacia la izquierda a medida que aumenta el nivel educativo.

3.2.3. Variable numéricas con comuna



Se observa que en el sur de la ciudad hay mayor cantidad de encuestados con niveles de inicial, primario y secundario completo, mientras que el norte (particularmente el barrio de Palermo) tiene mayor cantidad de personas con estudios superiores. En menor medida también las comunas del este (comúnmente llamado el çentro"de la ciudad) destacan por la cantidad de encuestados con nivel superior.

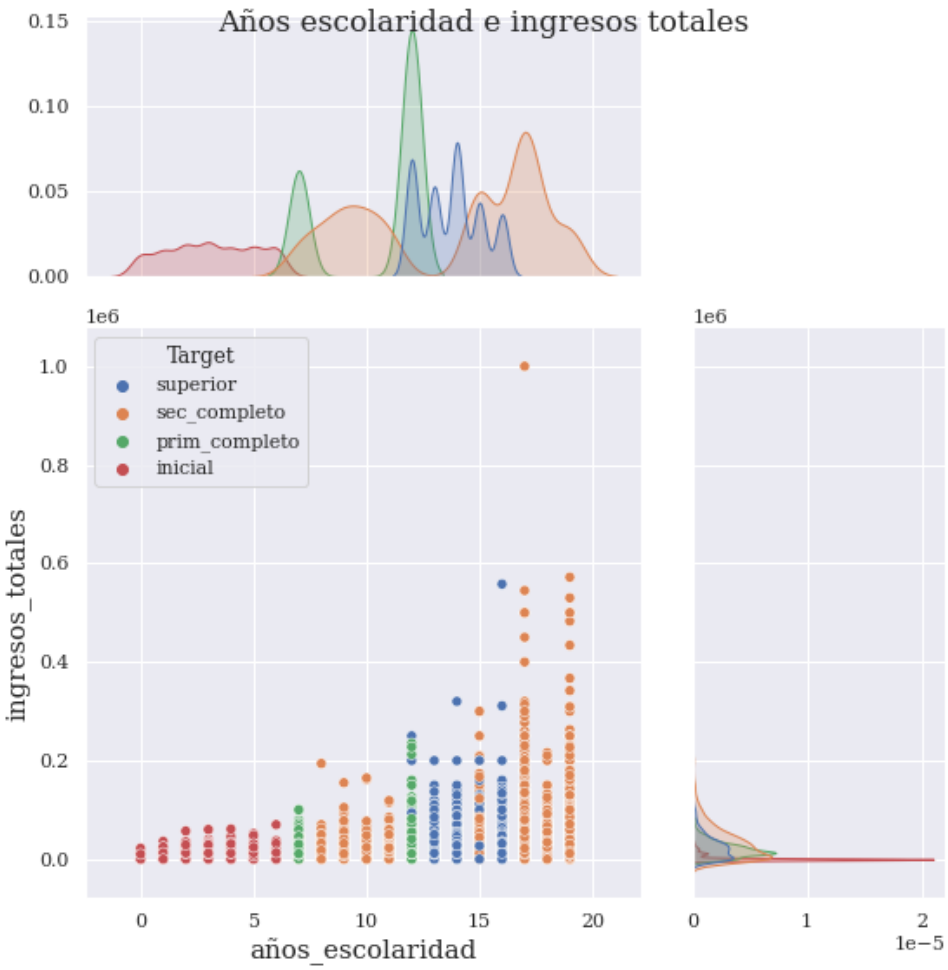


Lo que podemos observar en los últimos dos gráficos es una clara división geográfica del nivel educativo:

- Las comunas del norte son las que tienen mayor nivel educativo.
- Las comunas del centro tienen niveles medios.
- Las comunas del sur (con las comuna 6 en el centro de la ciudad como outlier) y la comuna 1 en el este son las que tienen niveles más bajos.

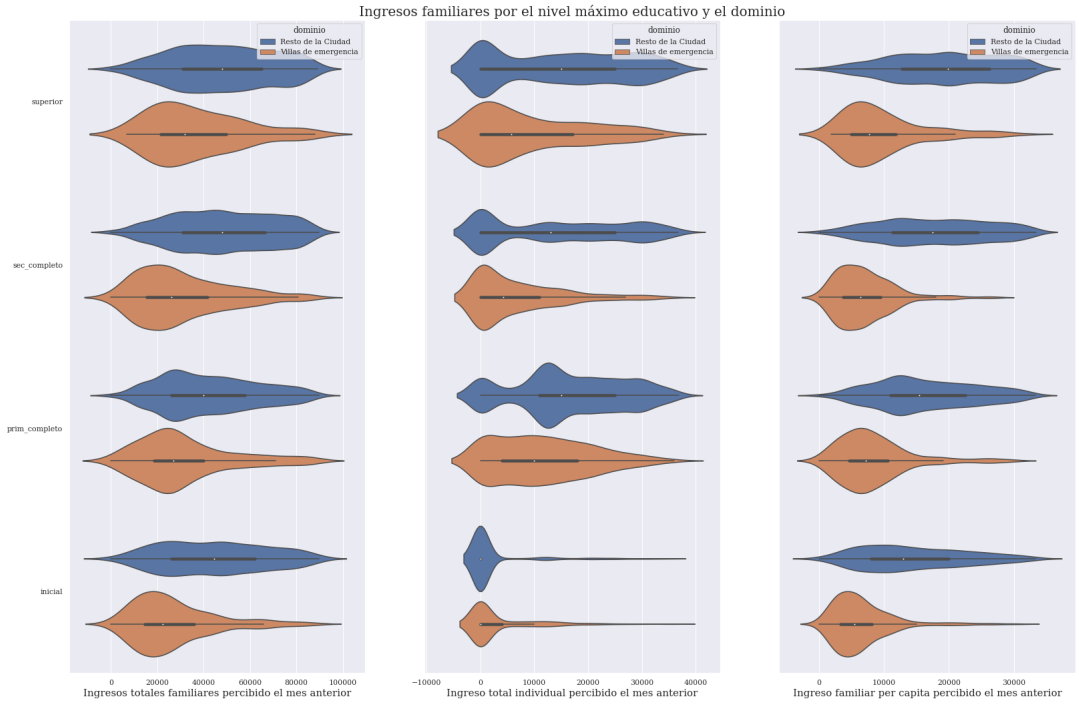
3.3. Análisis multivariado

Probamos de cruzar años de escolaridad, nivel máximo educativo y los ingresos totales.

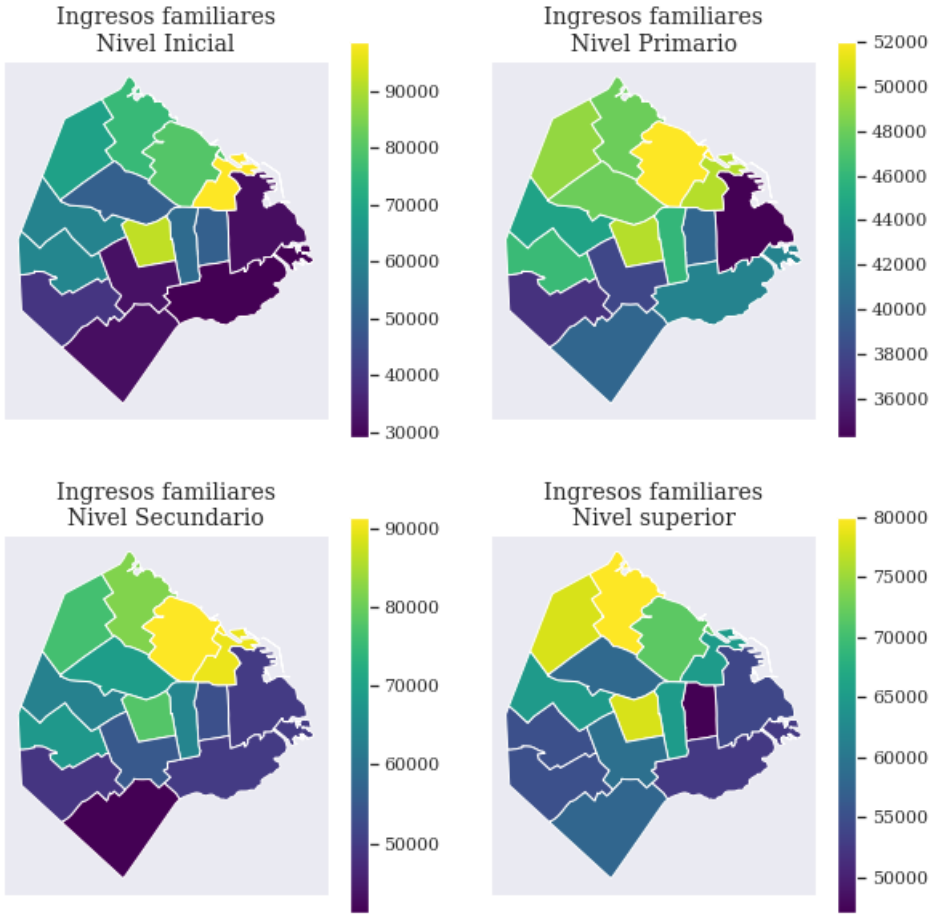


Conclusiones de la visualización:

- Hasta los 6 años, como era esperable, todos los casos llegan al nivel inicial.
- Vemos dos años en que aparece el primario completo: 7 y 12 años. Estimamos que se debe a la división entre los que comenzaron su educación en la primaria y los que comenzaron en el nivel inicial.
- A partir de los 12 años vemos un aumento consistente de los ingresos totales.



Aquí obtuvimos un descubrimiento interesante: no importa el nivel máximo educativo, los casos que no provienen de villas de emergencia (dominio="villas_de_emergencia") obtienen en promedio ingresos más altos en todos los niveles educativos. El alcanzar estudios superiores no parece homogeneizar ambos conjuntos. Esto se puede observar en el segundo gráfico, ya que el violín naranja acumula mayor cantidad de casos hacia la derecha, en comparación con los violines azules que tienen una mayor distribución.



Aquí podemos observar que a medida que avanza el nivel educativo máximo se atenúan levemente las diferencias de ingresos familiares entre comunas. Queda pendiente cruzar estos datos con la edad, para saber si el hecho de incluir a menores de edad está sesgando los valores para nivel inicial, primario y secundario.

4. Modelos analíticos

4.1. Tratados de nulos

4.1.1. Variables reemplazadas con moda

4.2. Target

4.2.1. Borrado de variables

4.3. División de train y test

4.3.1. Trian

One Hot Encoding

División de X e Y

4.3.2. Test

One Hot Encoding

División de X e Y

Comparación de cantidad de columnas

4.4. Buscando el mejor modelo